N° d'enregistrement

au C.N.R.S. AO 5178

ARTICLE PRINCIPAL

Recouvrant en partie

LA THESE DE DOCTORAT ES SCIENCES Mathématiques

presentée

A LA FACULTE DES SCIENCES

de PARIS

par

M. I.C. LERMAN

pour obtenir

le grade de Docteur Es. Sciences

Sujet de la Thèse : CONTRIBUTION AU PROBLEME DE LA VALIDITE DES RESULTATS D'UNE CLASSIFICATION AUTOMATIQUE.

soutenue le 18 janvier 1971

devant la Commission d'Examen

MM. D. DUGUE ..... Président

J.P. BENZECRI..... Examinateurs

R. DESCOMBES.

N° d'enregistrement

au C.N.R.S. AO 5178

ARTICLE PRINCIPAL

Recouvrant en partie

LA THESE DE DOCTORAT ES SCIENCES Mathématiques

presentée

A LA FACULTE DES SCIENCES

de PARIS

par

M. I.C. LERMAN

pour obtenir

le grade de Docteur Es. Sciences

Sujet de la Thèse : CONTRIBUTION AU PROBLEME DE LA VALIDITE DES RESULTATS D'UNE CLASSIFICATION AUTOMATIQUE.

soutenue le 18 janvier 1971

devant la Commission d'Examen

MM. D. DUGUE ..... Président

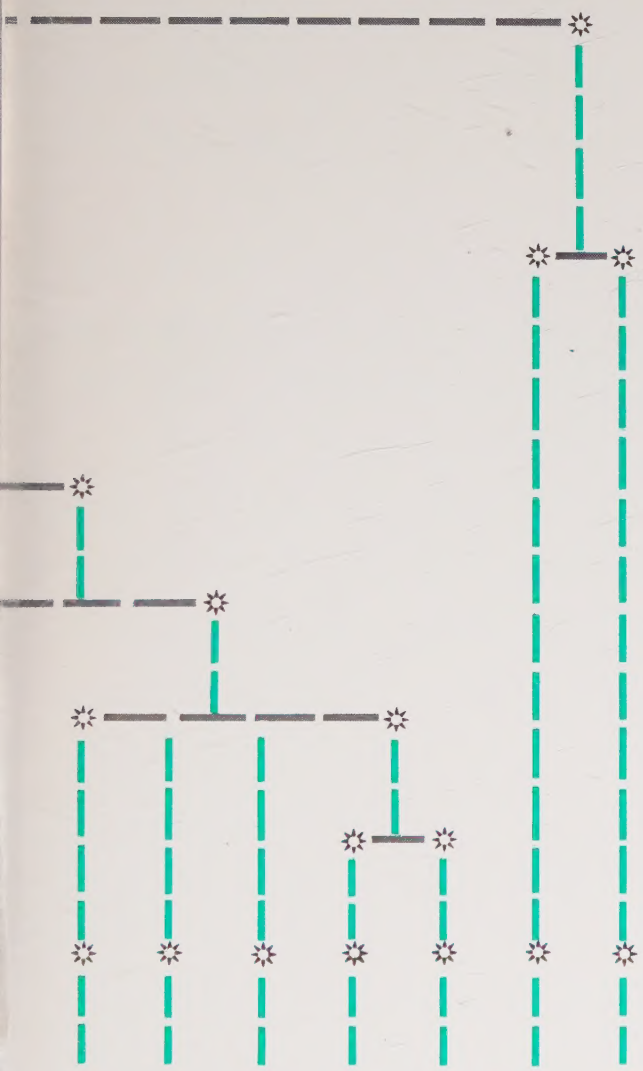
J.P. BENZECRI..... Examinateurs

R. DESCOMBES.

UNIV.
LEIDEN
BIBL.

LES BASES DE LA CLASSIFICATION AUTOMATIQUE

I. C. LERMAN



Collection « PROGRAMMATION »

GAUTHIER-VILLARS

Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41550026_0011

Les bases de la classification automatique

COLLECTION PROGRAMMATION
Publiée sous la direction de L. NOLIN

Publications de l'Institut de Programmation
de la Faculté des Sciences de Paris

Les bases de la classification automatique

par

I.C. LERMAN

GAUTHIER-VILLARS ÉDITEUR
55, quai des Grands-Augustins — PARIS 6e
1970

DANS LA MEME COLLECTION

Parus :

ARSAC J., LENTIN A., NIVAT M. et NOLIN L.— Algol. Théorie et pratique.

GROSS M. et LENTIN A.— Notions sur les grammaires formelles.

FAURE R.— Eléments de la recherche opérationnelle. Avec la collaboration de *N.S. Guillot Le Garff* et de *M. Bloch*.

NOLIN L.— Formalisation des notions de machine et de programme.

Sous presse :

FAURE R. et HEURGON E.— Structures ordonnées et algèbres de Boole.

© GAUTHIER-VILLARS, 1970

Toute reproduction, même partielle, de cet ouvrage est interdite. La copie ou reproduction, par quelque procédé que ce soit : photographie, microfilm, bande magnétique, disque ou autre, constitue une contrefaçon possible des peines prévues par la loi du 11 mars 1957 sur la protection des droits d'auteur.

TABLE DES MATIÈRES

Pages

INTRODUCTION & RESUME IX

Chapitre 0 — SUR QUELQUES QUESTIONS LIEES AU TREILLIS DES PARTITIONS 1

I — Treillis des partitions d'un ensemble fini 1

1 — Définitions et propriétés 1
2 — Dénombrement de chaînes de $\mathcal{R}$ 3

II — Partitions d'un entier n 4

1 — Généralités 4
2 — Représentations 5

III — Type d'une partition et cardinal du graphe de la relation d'équivalence correspondante 5

1 — Dénombrement de partitions de type donné 5
2 — Type d'une partition et cardinal du graphe correspondant 6

IV — Espaces ultramétriques 10

1 — Propriétés 10
2 — Chaîne du treillis des partitions d'un ensemble fini et espaces ultramétriques associés 11

Chapitre I — MESURE DE SIMILARITE ET PREORDONNANCE 13

I — Introduction 13

1 — Caractères et attributs 13
2 — Signification d'un attribut pour une classification 14
3 — Choix des attributs 14

II — Définition d'une mesure de similarité 15

1 — Généralités 15
2 — Définitions et propriétés 16

Pages

III — Préordonnance associée à une mesure de similarité 19

1 — Définitions et propriétés générales 19
2 — Ecart entre deux préordonnances associées à deux similarités . . 19

IV — Proposition de quelques préordonnances particulières sur E 24

1 — La préordonnance la "plus fine" 24
2 — La préordonnance la "moins fine" 25
3 — Préordonnances associées à des indices centrés et réduits 26

Chapitre II — CRITERES 27

I — Introduction 27

II — Critères définis à partir d'une matrice d'incidence des données 28

1 — Généralités 28
2 — Définition d'un critère 28

III — Critères définis à partir d'une relation binaire symétrique sur *E*. . . . 31

1 — Origine de la relation 31
2 — Représentation d'une partition 31
3 — Définition d'un critère 31
4 — Aspect limite du critère défini 32

IV — Critères définis à partir de la donnée d'une famille finie de partitions . 33

1 — Origine de la donnée 33
2 — Représentation d'une partition 34
3 — Définition d'un critère 34
4 — Objets équivalents 37

V — Critères définis à partir de la donnée d'un indice de similarité 38

A — *Recherche d'une partition* 38

1 — Introduction 38
2 — Critères définis pour la recherche d'une partition 38

B — *Recherche d'une chaîne de partitions* 39

1 — Représentation d'une chaîne de partitions 39
2 — Définition d'un critère 40

VI — Critères définis à partir d'une préordonnance sur E. 41

A — *Recherche d'une partition* 41

1 — Représentation d'une partition 41
2 — Définition d'un critère 41
3 — Comparaison des deux critères 43
4 — Distribution asymptotique de $|\omega \cap R \times S|$ 49
5 — Autres critères 51

Pages

B — *Recherche d'une chaîne de partitions* 54

1 — Représentation d'une chaîne de partitions 54
2 — Généralisation des critères précédents 54
3 — Rang lexicographique et nombre d'inversions 57

VII — Conclusion 63

Chapitre III — ALGORITHMES 65

I — Introduction 65

II — Algorithmes des transferts 65

1 — Définition 66
2 — Calcul des partitions centrales 66
3 — Application de l'algorithme à d'autres critères 67

III — Algorithme "lexicographique" 67

1 — Définition d'une préordonnance ultramétrique ω_u liée à ω . . . 68
2 — Algorithme permettant d'obtenir la préordonnance ultramétrique ω_u définie par H 68
3 — Hiérarchie de classifications associée à une ordonnance 71

Chapitre IV — ASPECTS STATISTIQUES 75

I — Introduction 75

II — Deux problèmes de convergence relatifs à la méthode des partitions centrales 75

A — *Premier problème* 75

1 — Introduction 75
2 — Proposition 75

B — *Deuxième problème* 77

1 — Introduction 77
2 — Proposition 1 78
3 — Proposition 2 78

III — H-classificabilité 80

1 — Introduction 80
2 — Ecartement entre la structure de ω et celle d'une préordonnance ultramétrique 81
3 — Test de l'hypothèse N de la non existence d'une h-classification "naturelle" 83

Annexe I, Programmes 89

Pages

Annexe II, Exemples réels 105

1 — Un exemple de géographie économique 105
2 — Un exemple de psychosociologie 107
3 — Un exemple sur l'évolution des espèces 110

BIBLIOGRAPHIE 115

INTRODUCTION ET RÉSUMÉ

La plus vaste classification que l'homme ait édifiée est celle du règne animal et du règne végétal. Bien que contestée dans ses ramifications, cette classification a été l'indispensable langage commun à tous les naturalistes et a constitué la matière brute dont l'élaboration suscite encore maintenant de nombreuses recherches fructueuses.

L. Blaringhem introduisant *Les Mathématiques de l'hérédité* de G. Malécot rappelle que "Francis Galton vers 1880, après Ad. Quetelet (1835), avant Alphonse Bertillon dont les *Instructions signalétiques* (1893) sont en usage familier, réalisa les portraits composites. Tous les Indiens, dit-il, ont, à première vue, des traits communs qui les distinguent des autres races ; les Anglais de même ne sont pas confondus avec les Espagnols ; et surtout, les membres d'une même famille ont dans leurs traits, dans leurs gestes, dans leur marche une allure qui malgré l'âge et le sexe donnent un air de famille", forment une classe "naturelle" dirons-nous ?

Utilisé couramment par les taxinomistes, le concept de classe "naturelle" reste mal défini. Danser (1950) s'aperçoit de la difficulté de donner un statut scientifique à ce terme. Comme déjà le faisait remarquer Vicq d'Azyr, (1792), la définition d'une classe par une propriété caractéristique, (classe monothétique), est trop rigide pour les Sciences de la nature et de l'homme. Il faut attendre Beckner (1959) pour une formulation intuitive, (qui laisse donc une part d'arbitraire), d'une classe "naturelle". Une telle classe se réfère à un ensemble fini d'attributs tel que chaque élément de la classe possède une proportion importante, mais non spécifiée, de ces attributs et réciproquement chaque attribut est possédé par une proportion importante des éléments de la classe, sans qu'il soit nécessaire qu'un des attributs soit possédé par tous les éléments de la classe. On appelle une telle classe polythétique (par opposition à monothétique).

C'est relativement à une visée classificatoire (Gilmour, 1937) que le spécialiste établit les attributs de description de la population qu'il étudie ; pour suivre la démarche d'Adanson (1757) : "Je me contenterai de rapprocher les objets suivant le plus grand nombre de degrés de leurs rapports et de leurs ressemblances... Les objets ainsi réunis formeront plusieurs petites familles que je réunirai encore ensemble, afin d'en faire un tout dont les parties soient unies et liées intimement".

Respectant l'intuition du spécialiste, il importe par conséquent pour une bonne classification (partition) de la population étudiée, que, relativement aux attributs de description, le degré de ressemblance de deux objets soit grand s'ils appartiennent à la même classe et petit sinon. De plus une hiérarchie de classifications emboîtées (chaîne ordonnée de partitions) respecte d'une manière satisfaisante les ressemblances entre objets, si la ressemblance de deux objets d'une même classe est d'autant plus grande que le niveau de la hiérarchie où ils se trouvent réunis pour la première fois est plus élevé (que la partition qui les réunit est plus fine).

Un des intérêts pratiques les plus immédiats de la taxinomie est de reconnaître un nouvel individu appartenant à une population déjà étudiée en lui attribuant son "adresse" dans la classification ; par

exemple savoir à quelle espèce appartient un poisson qu'on vient de découvrir. Il s'agit là d'un problème de *classement* qu'il ne faut nullement confondre avec celui de la *classification* où l'on souhaite dans ce cas, comme dit P. Dagnélie, "définir ou faire des classes".

Un objet plus élevé de la taxinomie est la connaissance d'un système (par exemple celui des vivants pour le règne animal). En cela le spécialiste peut être aidé par la détermination statistique, parmi les propriétés qui servent à décrire une population, de celles qui expliquent le mieux la classification obtenue selon le critère présenté ci-dessus.

Une telle synthèse de l'information apparaît comme un préalable très précieux pour l'étude de populations dans des domaines nouvellement explorés tels que la sociologie végétale, l'archéologie, la psychologie, la sociologie, l'économie, etc.

L'importance du temps nécessaire au spécialiste pour se familiariser avec l'objet de sa recherche montre l'intérêt d'une synthèse *automatique*, ayant recours à des algorithmes de calcul. Dans ce cadre, P. Dagnélie trouve surprenant le désintérêt des statisticiens pour les problèmes de la classification. Mais en général l'importance des données est telle que la masse des calculs à mettre en œuvre ne peut être conçue sans l'existence des calculatrices électroniques ; or il n'y a guère plus de dix ans, en France tout au moins, que le développement des ordinateurs permet d'entreprendre des calculs importants.

Se faisant l'écho de plusieurs chercheurs, l'ouvrage de R. Sokal et P.H.A. Sneath (1963) *The principles of numerical taxonomy* marquait un pas décisif dans l'introduction des méthodes automatiques. Le point de départ de chacune des méthodes proposées était le choix d'un indice ou mesure de similarité, fonction numérique sur $E \times E$, associant à chaque couple d'objets, (x, y), x et y appartenant à E, un nombre positif, $S(x, y)$, censé mesurer la ressemblance des deux objets et établi à partir des vecteurs logiques

$$\alpha(x) = (x_1, x_2, \ldots, x_l, \ldots, x_T)$$

et

$$\alpha(y) = (y_1, y_2, \ldots, y_l, \ldots, y_T)$$

où $x_l = 1$ si l'objet x possède l'attribut α_l et 0 sinon, (de même pour y), $\{\alpha_1, \alpha_2, \ldots, \alpha_T\}$ étant l'ensemble des attributs.

Au choix de la similarité succédait le choix d'un algorithme établissant une classification ou une hiérarchie de classifications susceptible de respecter les ressemblances entre objets à travers les nombres $S(x, y)$.

Il restait que seule l'intuition du spécialiste guidait le choix de la mesure de similarité et que les algorithmes proposés étaient des procédures d'approximation plus ou moins justifiées et souvent ambiguës ; néanmoins les différentes méthodes donnaient entre elles et celles des méthodes non automatiques des résultats très voisins.

Dés lors un souci de rigueur mathématique dans l'élaboration des méthodes s'imposait.

Dans un tout autre contexte, C.T. Zahn (1964) étudiait le problème de trouver une relation d'équivalence Π qui "approxime au mieux" une relation symétrique R sur un ensemble fini E, dans le sens suivant : "Minimiser le cardinal de la différence symétrique des graphes de R et de Π dans $E \times E$".

Ce problème illustre exactement la nature mathématique de celui de la classification automatique : la donnée de base est une structure de type σ sur l'ensemble E des objets qui est moins "riche" que celle, τ, que l'on désire (partition ou chaîne de partitions). La démarche générale consiste à injecter la classe Θ, des structures de type τ, dans celle, Σ, des structures de type σ. Il s'agit alors de déterminer celui ou ceux des éléments de la classe Θ qui sont au mieux compatibles avec la donnée de base. Pour cela on définit un *critère* ; un critère est, très généralement, une relation de préordre (qui peut être partielle) sur l'ensemble Θ. Pour cette relation τ est préférable à τ' ($\tau' < \tau$) si, en un sens qu'il y a lieu de préciser, τ est plus proche de la structure définie par la donnée de base que ne l'est τ'. Le critère est le plus souvent défini à partir de la donnée d'une métrique ou pseudo-métrique sur Σ, mais nous verrons des cas où il n'en est pas ainsi.

Chacun des caractères de description, $a_1, a_2, \ldots, a_h, \ldots, a_p$, présente un nombre fini de modalités et

détermine une partition de E. Celle relative au caractère a_h, notée S_h, réunit deux objets s'ils sont identiques du point de vue de a_h. S. Régnier (1965) juge d'autant plus préférable une partition X que

$$\frac{1}{p}\sum_{h=1}^{p} d(X, S_h) \qquad (1)$$

est le plus petit ; $d(X, S_h)$ désignant le cardinal de la différence symétrique des graphes de X et de S_h dans $E \times E$. Les "partitions centrales" sont celles qui réalisent le minimum de (1).

De plus, S. Régnier définit un algorithme dit des "Transferts" qui en un nombre de pas acceptable permet d'obtenir une partition qui réalise un optimum local de (1).

Cette méthode était un net progrès par rapport aux techniques utilisées. La démarche revenait par ailleurs à adopter comme mesure de similarité de deux objets x et y la proportion des caractères qui présentent la même modalité chez x et y, S_{xy} ; à plonger le cube des relations sur E, $\{0, 1\}^{E\times E}$, dans $R^{n\times n}$ et à chercher les points "partition" de ce cube les plus voisins du point de $R^{n\times n}$ $(S_{xy} : (x, y) \in E \times E)$.

Il restait que l'imperfection des données était telle qu'on ne pouvait pas considérer en général les différentes partitions S_h comme également discriminantes ou, ce qui revient au même, le choix de la similarité qui suppose également significatives de la ressemblance deux modalités quelconques, ne pouvait toujours être satisfaisant. De plus la méthode ne se généralisait pas au cas de la recherche d'une chaîne de partitions.

S'inspirant des travaux de R.N. Shepard (1962), J.P. Benzecri et W.F. de la Véga retenaient comme donnée, non plus les distances ou similarités $S(x, y)$, mais le système d'inégalités entre distances appelé ordonnance (1965), ordre total sur l'ensemble F des paires d'objets distincts de E. Conformément à cette appellation nous avons préfére distinguer ce cas de celui plus général et le plus fréquent, où la donnée est une *Préordonnance*, préordre total sur F.

Bien qu'associée au choix d'une distance ou d'une similarité, on pouvait espérer qu'une telle donnée fût plus stable. Sur une question de B. Jaulin, (1966), nous entreprîmes d'étudier l'influence du choix de la similarité sur la préordonnance associée. Moyennant une définition axiomatique, assez générale pour inclure la totalité des indices de similarités proposés, nous montrions que si le nombre d'attributs possédés par un même objet de E est invariable la préordonnance associée est la même quel que soit l'indice choisi ; plus généralement si ce nombre d'attributs a une variance faible, la préordonnance associée, lorsqu'on remplace un indice par un autre, varie peu au sens d'une métrique naturelle sur l'ensemble des préordres totaux sur F. Nous terminions en proposant des préordonnances qui nous semblaient naturelles, et qui ne dépendaient plus d'une seule similarité mais d'un couple ou d'un triplet. Cette étude fait l'objet du chapitre 1.

La donnée d'une partition P sur E définit sur F, un préordre à deux classes R et S ; où R (resp. S) est l'ensemble des paires réunies (resp. séparées) par la partition. Tout élément de R (paire de deux objets proches) précède tout élément de S (paire de deux objets éloignés).

Cherchant à adapter l'algorithme des "Transferts" au cas où la donnée est une ordonnance ω, J.P. Benzécri proposait de chercher la partition P qui "maximise le nombre d'inégalités spécifiées par P et compatibles avec ω". En d'autres termes, il s'agissait de maximiser $|\omega \cap R \times S|$, où ω désigne le graphe, dans $F \times F$, de la préordonnance et où (R, S) est associé à la partition P.

W.F. de la Véga préféra maximiser le cardinal du complémentaire dans $F \times F$ de la différence symétrique entre ω et $R \times S$, et envisagea le problème de la comparaison des deux critères en 1965. Ce problème, que nous reprîmes en 1967, nous a permis, avec des méthodes combinatoires, de montrer définitivement l'intérêt du critère de la différence symétrique.

Introduisant l'ensemble $\mathcal{R}_r$ des partitions pour lesquelles $|R| = r$ et Ω désignant l'ensemble de toutes les ordonnances sur E ; la distribution de $|\omega \cap R \times S|$, sur $\Omega \times \mathcal{R}_r$ muni d'une mesure uniforme est, pourvu que le cardinal de E ne soit pas trop petit, très approximativement *normale*. Sous ce point de vue, le critère de la Véga paraît être obtenu en centrant la

variable $|\omega \cap R \times S|$; nous proposons donc, comme nouveau critère, de réduire la variable centrée. Une telle statistique permettra en outre de tester la signification d'une classification obtenue par un algorithme.

Devant notre ignorance des exigences que doit satisfaire une classification, on est conduit parfois à limiter la confiance que l'on peut accorder à tout critère qui préordonne totalement l'ensemble des partitions. Conformément à la définition d'un ordre partiel sur l'ensemble des partitions, posée par W.F. de la Véga —définition qui utilise la notion de dominance statistique— J.P. Benzécri étudie un critère "qui ne donne pas toujours entre deux classifications une préférence ou une équivalence mais sur qui, lorsqu'il en donne, on est comme forcé de s'accorder". Ce critère dit d'ordre latéral est sûr mais partiel, son intérêt pratique reste à établir.

Nous avons bien établi deux critères d'ordre partiels qui peuvent être énoncés à propos de la recherche d'une classification mais qui sont plus naturels pour la recherche d'une hiérarchie de classifications, (chaîne de partitions), ce qui semble être l'idéal de ce que désirent les taxinomistes.

Bien que la nature profonde des techniques utilisées dans ce cas soit la même que pour la recherche d'une partition, ce problème ne se ramène pas mathématiquement au précédent. Introduisant la notion de préordonnance ultramétrique, dont la donnée est équivalente à celle d'une chaîne de partitions, nous avons généralisé les différents critères.

Nous avons établi dans ce cadre un algorithme qui est très lié à la nature de la donnée de base, préordonnance sur E. Cet algorithme que nous appelons "lexicographique" s'apparente à celui de P.H.A. Sneath (1957) ; son caractère très naturel nous a fait soupçonner une propriété d'optimalité, que nous avons cherché à expliciter. Ceci nous a conduit à définir le premier des deux critères d'ordre partiel, mentionnés ci-dessus, que nous avons désigné par "critère d'ordre lexicographique". Quant au second critère dit "du nombre d'inversions", il est plus classique, s'apparente au critère de la différence symétrique et tient son origine dans les travaux de M.G. Kendall. Nous avons tenu à juger le premier critère du point de vue du second.

Le chapitre 2 est consacré à la définition et à l'étude des critères selon que la donnée de base est :

a) le tableau des données $A \times E$ (où A est l'ensemble des attributs et E celui des objets) ;

b) une relation binaire symétrique sur E ;

c) une famille finie de partitions sur E ;

d) une distance ou pseudo-distance, (mesure de similarité), sur E ;

e) une préordonnance sur E.

Au chapitre 3 nous définissons d'une part l'algorithme des "transferts" ; d'autre part l'algorithme "lexicographique" et nous en étudions les propriétés.

La population E d'objets est en général un échantillon d'un ensemble E' plus vaste, qu'il est souvent impossible d'observer dans sa totalité. Par ailleurs une description exhaustive des éléments de E est encore souvent impossible ; les caractères ou attributs, retenus pour la description, constituent un échantillon d'un ensemble plus vaste de caractères ou attributs. Considérant cette situation, S. Régnier a étudié (1966) deux problèmes de convergence statistique qui permettaient de se rendre compte du comportement asymptotique des partitions centrales ; c'est ce que nous analysons dans la première partie du chapitre 4. Puis, dans le cadre où la donnée est une préordonnance ω, nous envisageons le problème de juger si, au vu des données, (ici de ω), il n'est pas *"artificiel"* d'imposer sur E une structure de chaîne de partitions, censée respecter les ressemblances entre objets. D'où la proposition d'une statistique susceptible de mesurer l'aptitude de E à être organisé en une hiérarchie de classifications, à être h-classifié dirons-nous ; et, la définition d'un test éprouvant l'hypothèse de la non existence d'une h-classification "naturelle" ; hypothèse que nous caractérisons ici d'une manière qui nous semble satisfaisante.

Enfin, pour terminer cette présentation, indiquons que le chapitre 0 introduit le treillis des partitions d'un ensemble fini et celui d'un entier, fixant ainsi le vocabulaire utilisé par la suite. Nous traitons dans ce chapitre de certaines questions de nature

combinatoire auxquelles nous nous sommes trouvés confrontés, notamment lors de la comparaison des critères. Ce chapitre reste essentiellement de référence.

Toutes les méthodes que nous considérons dans ce texte supposent donnée une information relative à la ressemblance des objets. Cette information est généralement établie sur les apparences sensibles, les caractères externes des objets, sans tenir compte des liaisons possibles entre ces caractères, c'est-à-dire de la structure sous-jacente qui parfois fait intervenir des informations extérieures aux objets (par exemple historiques, stratigraphiques etc.). Les classifications proposées sont donc phénotypiques. Relativement aux Sciences de la nature, on oppose à ce type de classifications celui des classifications phylogéniques, (par filiation), établies à partir de la parente présumée des espèces. Bien que leur thèse puisse prêter à discussion, R. Sokal et P. Sneath insistent dans leur ouvrage en assurant que le seul moyen d'établir une parenté est de comparer avec le plus de soin possible les manifestations externes [cf. pp. 55, 94, 216]. Il arrive même, disent-ils, que soit retenue une classification qui contredise ce que l'on suppose de la parenté des espèces. On croit que les crocodiles et les oiseaux ont des ancêtres communs dont ni lézard ni tortues ne descendent et pourtant on classe les crocodiles avec les autres reptiles, [cf. Sokal et Sneath p. 226]. Notre dernier mot rejoindra le premier ; la condensation de l'information, que fournissent nos classifications, permettra l'élaboration d'hypothèses et par conséquent de recherches futures.

Qu'il me soit permis de terminer en exprimant mes plus vifs remerciements à Messieurs B. Jaulin et S. Régnier pour les nombreuses conversations que j'ai pu avoir avec eux et dont chacune enrichissait mes associations d'idées. Tous mes remerciements à Monsieur M. Eytan qui a contribué à la rédaction du chapitre 0 et nous a fourni une caractérisation d'une matrice de distances ultramétrique.

Je dois au Professeur J.P. Benzécri des réactions et commentaires qui m'ont été d'un grand profit.

ERRATUM

Les symboles de référence indiquent successivement la page, la colonne et la ligne ; ainsi (p. 15, d, l.-6) désigne page 15, colonne de droite, ligne 6 à partir du bas.

(p. 11, g, l.16) au lieu de : théorème 2, lire : théorème 1

(p. 12, d, l.5 à 6) le segment joignant $\{g\}$ à $\{h, i, j\}$ dans le diagramme doit être éffacé et remplacé par un segment joignant $\{g\}$ à $\{d, e, f, g\}$

(p. 15, d, l.-6) au lieu de (resp. $x_i y_i = 0$), lire : (resp. $x_i = y_i = 0$).

(p. 17, d, l.13) au lieu de : l'axiome 3, lire : l'axiome 2.

(p. 19, g, l.-12) au lieu de : $\{b, d\} < \{c, d\}$, lire $\{b, d\} = \{c, d\}$.

(p. 20, g, l. -5) ajouter à la suite $(\{d, f\}, \{b, c\})$, $(\{b, c\}, \{d, f\})$.

(p. 24, dans le diagramme) ôter le 1 de la position (6,6) et le mettre en position (4,6) ; de la sorte $\nu = 2+4+1+2 = 9$; l'axe vertical correspond à u+v.

(p. 25, g, l.-4) au lieu de : $\{b, c\} < \{d, e\}$, lire : $\{b, e\} < \{d, e\}$.

(p. 30, d, l. 10 et 11) il faut lire 3+1+1+0+1+2+1+1+2+1+1+2+2+2+1+1+1+1= 24

(p. 32, g, l.-1) dans la case (b, c) mettre une croix en même temps que le point

(p. 39, g, l.9) au lieu de : proposition 10 (Chap. 0 § 4), lire : propositions 1 et 2 (Chap. 0 § IV.2)

(p. 40, g, l. 4) au lieu de : b) et c), lire : a) et b).

(p. 49, g, l. 21) au lieu de : $|\omega \cap R \times R|$, lire : $|\omega \cap R \times S|$.

(p. 56, d, l. 17) il faut lire : 7+2+4+0+11+11+4+4+4+2+2+2+0 = 55

(p. 56, d, l.-1) au lieu de : analogue à 2), lire : analogue à 2'.2).

(p. 63, d, l.-10) la référence doit être §VI.A.5

(p. 69, g, l.13) il faut lire : 1 2 2 3 4 4 4 4 4 4 4 4 4 4 4

(p. 79, g, l.8) au lieu de : $\Pr\left\{X \in \bigcap_{p\ k} C_p\right\} = 1$, lire $\Pr\left\{X \in \bigcup_{p\ k} C_p\right\} = 1$.

CHAPITRE 0

SUR QUELQUES QUESTIONS LIÉES A L'ENSEMBLE DES PARTITIONS D'UN ENSEMBLE FINI

I — *TREILLIS DES PARTITIONS D'UN ENSEMBLE FINI*

1 — Définitions et propriétés

Soit E un ensemble fini à n éléments.

Une *partition* de E est un ensemble de parties de E mutuellement disjointes dont la réunion est E. Ces parties sont les *classes* de la partition. Ainsi $\{a, b, c, d\}, \{e, f\}$ et $\{g\}$ sont les classes de la partition $\{\{a, b, c, d\}, \{e, f\}, \{g\}\}$ de

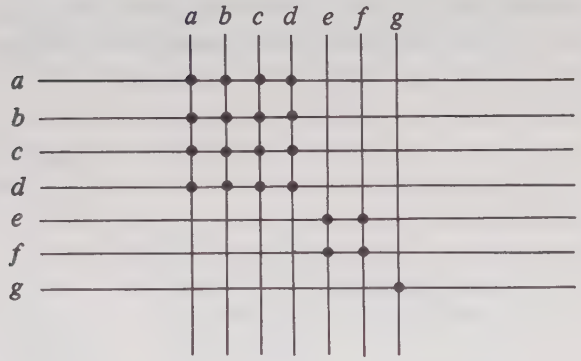
$$E = \{a, b, c, d, e, f, g\}$$

Désignons par $\mathfrak{P}(E)$ l'ensemble des partitions de l'ensemble E. Se donner une partition P de E revient à se donner une relation d'équivalence dans E (à savoir deux éléments x et y de E sont équivalents si et seulement si tous deux appartiennent à une même classe de P) que nous noterons encore P, le contexte étant en général suffisant pour éviter toute ambiguïté.

Le graphe de la relation d'équivalence P sera noté $\mathrm{Gr}(P)$;

$$\mathrm{Gr}(P) = \{(x, y) \mid x \in E, y \in E \quad \text{et} \quad xPy\}$$

$\mathrm{Gr}(P)$ est une partie de $E \times E$ que nous allons représenter dans le cas de l'exemple ci-dessus.



L'inclusion dans $E \times E$ permet de définir *un ordre*, noté $\leqslant$, sur $\mathfrak{P}(E)$; $P \leqslant P'$ signifiera $\mathrm{Gr}(P) \subset \mathrm{Gr}(P')$, ou encore $(\forall x, y \in E) : xPy \Longrightarrow xP'y$.

"La partition P est un raffinement de la partition P'" ; on dit que P *est plus fine que* P'. Ainsi la partition $\{\{a, b, c, d\}, \{e, f\}, \{g\}\}$ est plus fine que celle $\{\{a, b, c, d\}, \{e, f, g\}\}$ de $E = \{a, b, c, d, e, f, g\}$.

Dès lors l'ensemble ordonné $\mathfrak{P}(E)$ est muni d'une structure de treillis ; c'est-à-dire que toute paire de partitions $\{P, P'\}$, admet un plus grand minorant commun $P \wedge P'$ et un plus petit majorant commun $P \vee P'$.

La partition $P \wedge P'$ peut être définie par son graphe de la façon suivante : $\mathrm{Gr}(P \wedge P') = \mathrm{Gr}(P) \cap \mathrm{Gr}(P')$; en d'autres termes $x(P \wedge P')y$ si et seulement si xPy et $xP'y$.

$P \vee P'$ est défini par : $\mathrm{Gr}(P \vee P')$ est le plus petit graphe de relation d'équivalence contenant l'ensemble $\mathrm{Gr}(P) \cup \mathrm{Gr}(P')$, $\mathrm{Gr}(P \vee P')$ est le graphe de la fermeture transitive de la relation "P ou P'", soit encore $x(P \vee P')y$ si et seulement s'il existe une suite $z_0, z_1, \ldots, z_k$ d'éléments de E vérifiant : pour tout $i = 0, 1, \ldots, k-1$, $z_i P z_{i+1}$ ou $z_i P' z_{i+1}$ $z_0 = y$ et $z_k = y$.

Exemple : $P = \{\{a, b, c, d\}, \{e, f\}, \{g\}\}$,

$$P' = \{\{a, b\}, \{c, d\}, \{e, f, g\}\}$$

$$P \wedge P' = \{\{a, b\}, \{c, d\}, \{e, f\}, \{g\}\} \quad \text{et}$$

$$P \vee P' = \{\{a, b, c, d\}, \{e, f, g\}\}$$

Il est facile de voir que le treillis $\mathfrak{P}(E)$ ne dépend que de l'entier n, cardinal de E, nous le noterons ci-dessous par $\mathfrak{P}$. Le plus petit élément de $\mathfrak{P}$ est la partition la plus fine, celle dont chaque classe contient exactement un seul élément de E. Le plus grand élément de $\mathfrak{P}$ est la partition la moins fine, celle qui comporte une seule classe, E. La partition la plus fine est dite *discrète*, celle, la moins fine, *grossière*.

Rappelons quelques définitions concernant le treillis $\mathfrak{P}$.

a) Une partition P' couvre une partition P si on a

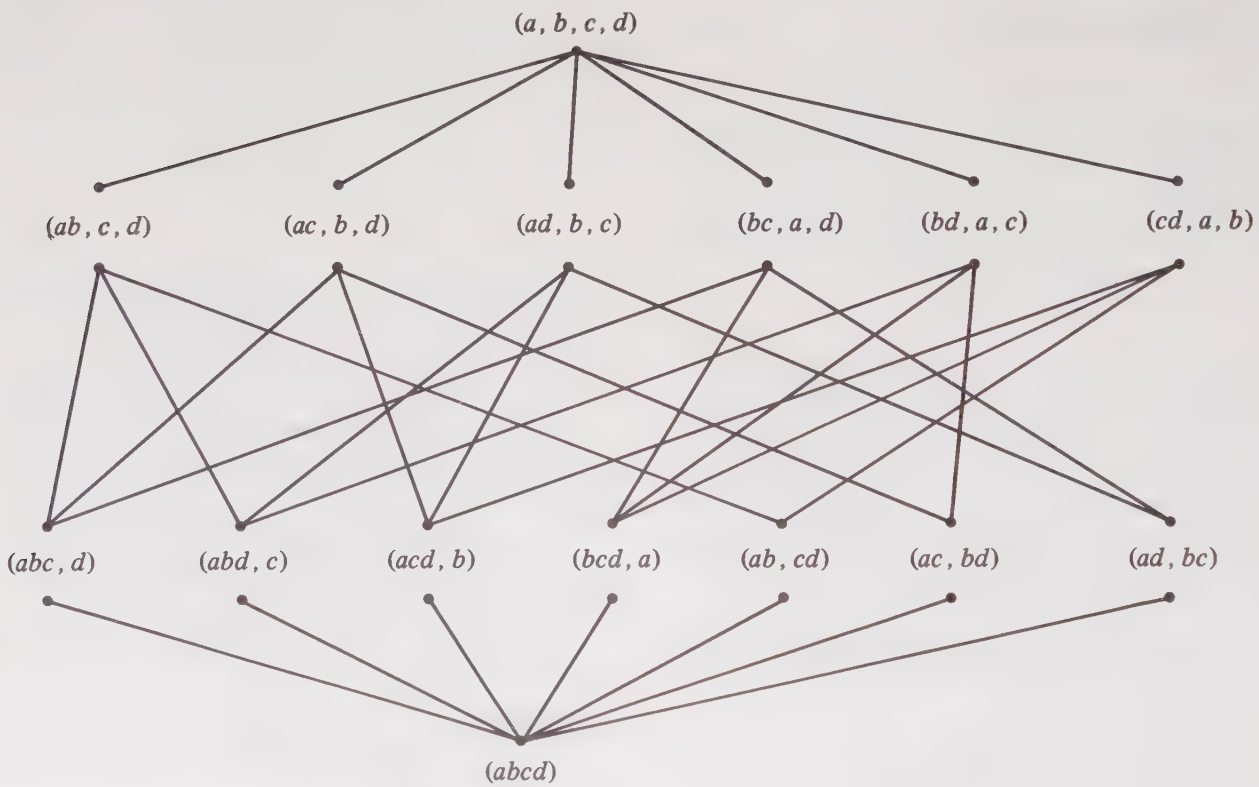
i) $P \leqslant P'$

ii) $\{Q \in \mathfrak{P} \mid P < Q < P'\} =]P, P'[= \emptyset$;

en d'autres termes, il n'existe pas de partition Q qui soit strictement entre P et P'.

Il est immédiat de voir que P' couvre P si et seulement si P' s'obtient en réunissant deux classes de P.

b) Les *atomes* de $\mathfrak{P}$ sont les éléments de $\mathfrak{P}$ qui couvrent la partition discrète P_0. $\mathfrak{P}$ est un *treillis atomique*, c'est-à-dire que tout P de $\mathfrak{P}$ s'écrit



$$P = \bigvee_{i=1}^{p} A_i$$

tous les A_i étant des atomes de $\mathfrak{P}$, ce sont les équivalences qui identifient deux éléments.

c) Une chaîne du treillis $\mathfrak{P}$ est un sous-ensemble totalement ordonné de $\mathfrak{P}$, $C = (P_1, P_2, \ldots, P_k)$ avec $P_1 < P_2 < \ldots < P_k$ et $P_i \in \mathfrak{P}$. La longueur d'une chaîne est égale au nombre d'éléments qu'elle contient moins un.

La hauteur d'une partition P, $h(P)$, est le maximum des longueurs de chaînes reliant la partition discrète, P_0 à P.

Il est facile de voir que si $c(P)$ est le nombre de classes de la partition $P \in \mathfrak{P}$; alors la hauteur est donnée par $h(P) = n - c(P)$.

La hauteur du treillis $\mathfrak{P}$ est la hauteur du plus grand élément, ici de la partition grossière pour laquelle $c(P) = 1$, soit $h(P) = n - 1$.

d) $\mathfrak{P}$ est un *treillis gradué* par la hauteur h ; c'est-à-dire :

i) $P' > P \Longrightarrow h(P') > h(P)$

ii) si P' couvre P, alors $h(P') = h(P) + 1$.

Dans ces conditions $\mathfrak{P}$ vérifie la *condition de Jordan-Dedekind,* à savoir : toutes les chaînes maximales entre deux éléments quelconques du treillis $\mathfrak{P}$ ont une même longueur (finie). En effet la longueur d'une chaîne maximale quelconque entre P et P' $(P' > P)$ est $h(P') - h(P)$.

Un treillis gradué de hauteur finie est *semi-modulaire* si l'on a :

$$h(P) + h(P') \geqslant h(P \vee P') + h(P \wedge P') ;$$

c'est le cas pour le treillis $\mathfrak{P}$.

Exemple : soit le treillis $\mathfrak{P}$ d'un ensemble à 4 éléments

$$E = \{a, b, c, d\}$$

on notera les partitions sous la forme : (abd, c) pour $\{\{a, b, d\}, \{c\}\}$.

Le treillis $\mathfrak{P}_4$ est figuré ci-contre par un diagramme de finesse.

2 — Dénombrement de chaînes de $\mathfrak{P}$

Nous nous intéressons seulement à l'ensemble $\mathfrak{C}(E)$ des chaînes de $\mathfrak{P}$ dont le premier élément est la partition discrète ; $C \in \mathfrak{C}(E)$ si

$$C = (P_0, P_1, \ldots, P_k)$$

avec $P_0 < P_1 < \cdots < P_k$ et où P_0 est la partition la plus fine.

Une chaîne $C \in \mathfrak{C}(E)$ sera dite *élémentaire* si pour tout i, P_{i+1} couvre P_i. Une chaîne $C \in \mathfrak{C}(E)$ sera dite *plus fine* que $C' \in \mathfrak{C}(E)$ si $C \supset C'$ i.e.

$$P_i \in C' \Longrightarrow P_i \in C$$

pour tout i. L'ensemble $\mathfrak{C}(E)$ muni de cet ordre est même un demi-treillis pour l'opération d'intersection avec pour plus petit élément $C_0 = \{P_0\}$. Dans ce demi-treillis les chaînes élémentaires sont des éléments maximaux, de longueur $n = |E|$.

Le cardinal de $\mathfrak{C}(E)$ dépend seulement de n, cardinal de E, notons-le $\psi(n)$.

Proposition 1 : On a : $\psi(n) = \sum_{k=1}^{n-1} S(n, k)\, \psi(k)$

où $S(n, k)$ désigne le cardinal de l'ensemble des partitions de n objets en k classes ; nombre de Stirling de seconde espèce.

Posons $\mathfrak{C}_k(E) = \{C \in \mathfrak{C}(E) \mid c(P_1) = k\}$ où $c(P_1)$ est le nombre de classes de la partition P_1. Les $\mathfrak{C}_k(E)$, $k = 1, 2, \ldots, (n-1)$, forment manifestement une partition de $\mathfrak{C}(E)$. On va s'attacher à déterminer le cardinal de $\mathfrak{C}_k(E)$. Pour un élément P_1 fixé comportant k classes, il y a $\psi(k)$ chaînes dans $\mathfrak{C}(E)$ pour lesquelles la partition de niveau 1 est P_1, c'est-à-dire autant que de chaînes de partitions sur l'ensemble des classes de P_1. Il y a $S(n, k)$ partitions différentes de n objets en k classes ; donc,

$$|\mathfrak{C}_k(E)| = S(n, k)\, \psi(k)$$

D'où :

$$|\mathfrak{C}(E)| = \psi(n) = \sum_{k=1}^{(n-1)} S(n, k)\, \psi(E)$$

Exemple : $n = 4$

$$\psi(4) = S(4,1)\,\psi(1) + S(4,2)\,\psi(2) + S(4,3)\,\psi(3)$$
$$= 1 \times 1 + 7 \times 1 + 6 \times 4$$
$$= 32$$

ce qu'on vérifiera directement sur le treillis $\mathfrak{P}_4$ figuré ci-dessus.

Proposition 2 : soit $\varphi(k)$ le nombre de chaînes élémentaires de longueur k. On a

$$\varphi(n) = (n-1)!\,n!/2^{n-1}$$

Une chaîne élémentaire C de $\mathfrak{C}(E)$ comporte exactement n éléments $C = (P_0, P_1, \ldots, P_{n-1})$, la partition P_k étant au niveau k et ayant donc $n-k$ classes. Si on fixe le k premiers termes $P_0, P_1, \ldots, P_{k-1}$ de C, il y a $\binom{n-k+1}{2}$ partitions P_k qui peuvent couvrir P_{k-1}, puisque P_k s'obtient par réunion de deux classes de P_{k-1}. Donc

$$\varphi(n) = \binom{n}{2} \cdot \binom{n-1}{2} \cdots \binom{2}{2}$$

d'où la formule.

Exemple : $n = 4, \varphi(4) = 3!\;4!/2^3 = 18$, ce qu'on vérifiera directement sur le diagramme ci-dessus.

II — *PARTITIONS D'UN ENTIER n*

1 — Généralités

Le *type* d'une partition P est la suite décroissante des cardinaux n_j de ses classes. L'*ordre* de la partition est la suite $(m_1, m_2, \ldots, m_n)$ où m_i est le nombre de classes de cardinal i. On a évidemment :

$$\sum_j n_j = n\,, \qquad \sum_i i m_i = n$$

Soit n un entier naturel donné. Une *partition de l'entier n* est la donnée d'un ensemble fini

$$\{n_1, n_2, \ldots n_k\}$$

d'entiers naturels strictement positifs vérifiant

$$n_1 + n_2 + \cdots + n_k = n.$$

Le fait qu'on ait pris les parts, i.e. les n_i, comme éléments d'un ensemble signifie que l'ordre n'est pas spécifié. On dit parfois que c'est une partition non ordonnée de n. On remarquera qu'une partition ordonnée (les n_i forment une suite) s'obtient en écrivant $n = 1 + 1 + + \cdots + 1$ et en plaçant $k-1$ cloisons dans les $n-1$ places disponibles. Ceci peut être fait de $\binom{n-1}{k-1}$ façons ; il y a donc $\binom{n-1}{k-1}$ telles partitions de n en k parts.

Pour en revenir aux partitions non ordonnées, on peut toujours se ramener au cas d'une suite décroissante en supposant $n_1 \geqslant n_2 \geqslant \cdots \geqslant n_k \geqslant 1$. Finalement donc une partition de l'entier n sera la suite décroissante $(n_1, n_2, \ldots)$ vérifiant

$$n_1 + n_2 + \cdots = n\,,$$

les parts étant les n_i.

On prendra garde de ne pas confondre partition d'un ensemble et partition d'un entier ; d'ailleurs certains désignent cette dernière notion par "partage d'un entier".

Si toutefois la terminologie utilisée a été préférée c'est (raisons historiques mises à part) qu'il y a un lien entre les deux notions. Munissons en effet l'ensemble des partitions de l'entier n d'une relation d'ordre $\leqslant$ définie comme suit, $(n'_j) \leqslant (n_i)$ si et seulement si il existe un i tel que

$$n_i = n'_{j_1} + n'_{j_2} + \cdots + n'_{j_k}\,.$$

Désormais Part (n) désignera l'ensemble des partitions de l'entier n muni de l'ordre $\leqslant$ ainsi défini.

Proposition 1 : Part (n) est le quotient de l'ensemble ordonné $\mathfrak{P}$ des partitions d'un ensemble de cardinal n par la relation d'équivalence "avoir même type que".

L'application qui à une partition $P \in \mathfrak{P}$ associe son type est un morphisme d'ensembles ordonnés de $\mathfrak{P}$ dans Part (n), qui est surjectif. Si P à k classes, n est partitionné en k parts.

Proposition 2 : soit k_i le nombre de fois où la part i ($i = n, \ldots, 2, 1$) intervient dans une partition de l'entier n. A cette partition de n on peut associer $\dfrac{n!}{\prod\limits_{1 \leqslant i \leqslant n} (i!)^{k_i} k_i!}$ partition de l'ensemble E de cardinal n.

Cherchons le nombre ν de partitions de l'ensemble E d'ordre $(k_1, k_2, \ldots, k_n)$. En permutant les objets d'une classe puis les classes, on obtient toutes les permutations de E, soit $n!$. Or on peut permuter les objets d'une classe de $i!$ façons, il y a k_i classes de cardinal i et il y a $k_i!$ façons de permuter entre elles les classes de même cardinal. Si bien que $\nu \times \prod\limits_{1 \leqslant i \leqslant n} (i!)^{k_i} k_i! = n!$.

L'ensemble ordonné Part (n) n'est pas un treillis, deux partitions de l'entier n n'ayant pas obligatoirement un infimum. Le morphisme défini à la proposition 1 n'est donc pas un morphisme de treillis, et en fait ne conserve pas l'opération V du treillis $\mathfrak{P}$. Cependant dans Part (n) toutes les chaînes maximales entre deux éléments ont même longueur.

2 — Représentations

Outre la représentation comme quotient de partitions d'un ensemble, les partitions d'un entier n admettent deux représentations intéressantes.

2.1 — *Graphe de Ferrers*

Soit $n_1 + n_2 + \cdots + n_k = n$ une partition de l'entier n en k parts (vérifiant comme convenu

$$n_1 \geqslant n_2 \geqslant \cdots \geqslant n_k \geqslant 1).$$

On peut en donner la représentation graphique suivante, appelée *graphe de Ferrers* :

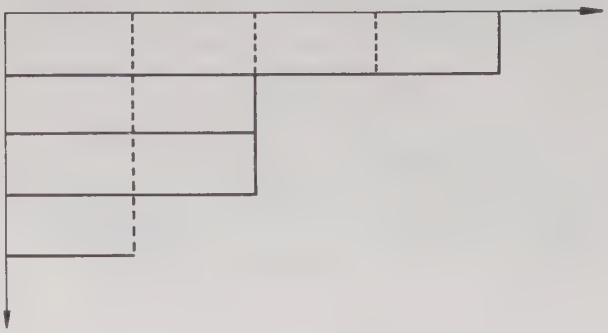
sur k lignes horizontales on porte des points, la $i^{\text{ème}}$ ligne portant n_i points.

Exemple : $4 + 2 + 2 + 1$ est représenté par le graphe de Ferrers

. . . .
. .
. .
.

Une variante de cette représentation (utilisée par G. Kreweras qui l'appelle alors "profil") consiste à introduire des axes de coordonnées et de porter des cases au lieu de points.

L'exemple ci-dessus a pour profil



Manifestement il n'y a pas de différence essentielle entre graphe de Ferrers et profil ; nous nous en tiendrons au premier.

La représentation par graphe de Ferrers visualise un fait important ; si on fait pivoter un graphe de Ferrers d'une partition de l'entier n d'un angle droit, on obtient un nouveau graphe de Ferrers, correspondant à une autre partition du même entier n, dite partition conjuguée. Ainsi dans l'exemple ci-dessus on obtient le graphe de Ferrers de la partition $4 + 3 + 1 + 1$ du même entier 9.

. . . .
. . .
. .
.

III — *TYPE D'UNE PARTITION ET CARDINAL DU GRAPHE DE LA RELATION D'EQUIVALENCE CORRESPONDANTE*

1 — Dénombrement de partitions de type donné

Soit f une surjection d'un ensemble E de cardinal n sur l'ensemble $\{1, 2, \ldots, k\}$. La partition définie

par l'application f est telle que les classes E_i de cette partition vérifient $E_i = f^{-1}(i)$. On dit que f détermine une partition aux *classes étiquetées*. Si n_i est le cardinal de E_i, le type de la partition est $(n_1, n_2, \ldots, n_k)$. Désormais on désignera par $\mathfrak{F}$ l'ensemble des partitions aux classes étiquetées de type $(n_1, \ldots, n_k)$, et par Λ l'ensemble des partitions de type $(n_1, \ldots, n_k)$. On remarque que

$$|\mathfrak{F}| = \frac{n!}{n_1!\, n_2! \ldots n_k!}$$

Partageons la suite des cardinaux, $(n_1, n_2, \ldots, n_k)$, en sous-suites. Une même sous-suite est formée de cardinaux tous égaux entre eux. Si les effectifs de ces dernières sous-suites sont respectivement

$$k_1, k_2, \ldots, k_r ;$$

on a

$$|\Lambda| = \frac{n!}{k_1!\, k_2! \ldots k_r!\, n_1!\, n_2! \ldots n_i! \ldots n_k!}$$

2 — Type d'une partition P de E et cardinal du graphe Gr (P) correspondant

Désignant par F l'ensemble $P_2(E)$ des parties à deux éléments de E, la donnée d'une partition P de l'ensemble E en k classes $E_1, E_2, \ldots, E_k \in E/P$ détermine une partition de F en deux classes :

$$R(P) = \{\{x, y\} \in F : \exists j, x \in E_j \quad \text{et} \quad y \in E_j\}$$

$$S(P) = F - R(P) =$$

$$\{\{x, y\} \in F : \exists i \neq j, x \in E_i \quad \text{et} \quad y \in E_j\}$$

La donnée de $G(P)$ est équivalente à celle de $R(P)$ et on a $|\mathrm{Gr}(P)| = 2|R(P)| + n$. On se ramène donc à étudier les rapports entre le type d'une partition P et le cardinal de l'ensemble R correspondant.

Proposition 1 : si $(n_1, n_2, \ldots)$ est le type de la partition P l'ensemble, on a :

$$|R(P)| = \sum_i n_i (n_i - 1)/2$$

$$|S(P)| = \sum_{i<j} n_i n_j \; ; \; |R(P)| + |S(P)| = n(n-1)/2$$

où $|\bullet|$ désigne le cardinal de $\bullet$.

Posons $R_i = \{\{x, y\} \in F : x \in E_i \text{ et } y \in E_i\}$; les R_i constituent une partition de $R(P)$ et

$$|R_i| = n_i (n_i - 1)/2 ,$$

d'où la première formule.

Posons $S_{ij} = \{\{x, y\} \in F : x \in E_i \text{ et } y \in E_j\}$; les S_{ij}, avec $i < j$, constituent une partition de $S(P)$ et $|S_{ij}| = n_i n_j$, d'où la deuxième formule.

Enfin la troisième formule résulte de ce que $R(P)$ et $S(P)$ constituent une partition de F, avec

$$|F| = n(n-1)/2 .$$

Proposition 2 : parmi toutes les partitions de E en k classes, celles qui réalisent le minimum de $|R|$ sont du type

$$(q+1, q+1, \ldots, q+1, q, q, \ldots, q, 0, \ldots)$$

où q est le quotient de la division de n par k : $n = kq + r$, $r < k$.

La démonstration se fait par récurrence sur n, en remarquant que

$$|R(P)| = \sum_i n_i (n_i - 1)/2 = \frac{1}{2}\left(\sum_i n_i^2 - n\right)$$

et donc qu'un minimum pour $|R(P)|$ sera atteint lorsqu'on aura un minimum pour $\sum_i n_i^2$.

Pour $n = k$ la propriété est triviale. On peut donc supposer la propriété vraie pour un certain $n > k$

$$r(q+1)^2 + (k-r)q^2 \leqslant \sum_i n_i^2, \quad \sum_i n_i = n .$$

Montrons que

$$(r+1)(q+1)^2 + (k-r-1)q^2 \leqslant \sum_i m_i^2 ,$$

$$\sum_i m_i = n + 1$$

ou encore que

$$r(q+1)^2 + (k-r)q^2 + 2q + 1 \leqslant (m_1 - 1)^2 + \sum_{i \geqslant 1} m_i^2 + 2m_1 - 1 .$$

Or m_1 est le plus grand des m_i, donc $m_1 \geqslant q+1$ et $2q+1 \leqslant 2m_1 - 1$. Par ailleurs $(m_1 - 1, m_2, \ldots, m_k)$ est une partition de l'entier n en k parts. L'hypothèse de récurrence implique donc que

$$r(q+1)^2 + (k-r)q^2 \leqslant (m_1 - 1)^2 + \sum_{i \geqslant 2} m_i^2 .$$

Il suffit d'ajouter membre à membre les deux dernières inégalités pour terminer la démonstration.

Proposition 3 : parmi toutes les partitions de E en k classes, celles qui réalisent le maximum de $|R|$ sont du type $(n-k+1, 1, 1, \ldots, 1, 0, \ldots)$.

La démonstration, comme la précédente procède par récurrence sur n et porte sur $\sum_i n_i^2$.

La propriété est triviale pour $n = k$. Supposons donc que pour un $n > k$ on a

$$(n-k+1)^2 + k - 1 \geqslant \sum_i n_i^2 \quad , \quad \sum_i n_i = n$$

et montrons que

$$(n+1-k+1)^2 + (k-1) \geqslant \sum_i m_i^2 \quad ,$$

$$\sum_i m_i = n+1$$

ou encore que

$$(n-k+1)^2 + 2(n-k+1) + 1 + k - 1 \geqslant (m_1 - 1)^2 + \sum_{i \geqslant 2} m_i^2 + 2(m_1 - 1) + 1.$$

Or $m_1 \leqslant (n+1) - k + 1 = n - k + 2$ donc

$$2(n-k+1) + 1 \geqslant 2(m_1 - 1) + 1$$

Par ailleurs $(m_1 - 1, m_2, \ldots, m_k)$ est une partition de l'entier n en k parts et donc par l'hypothèse de récurrence

$$(n-k+1)^2 + k - 1 \geqslant (m_1 - 1)^2 + \sum_{i \geqslant 2} m_i^2$$

Il suffit d'ajouter membre à membre les deux inégalités pour terminer la démonstration.

Proposition 4 : parmi toutes les partitions de E, celles qui réalisent le maximum de $|R(P) \times S(P)|$ sont telles que $|R(P)| = |S(P)|$.

En remarquant que $|R \times S| = |R| \times |S|$ et que $|R| + |S| = n(n-1)/2$ la proposition devient triviale. Une borne supérieure de $|R \times S|$ est donc $(n(n-1)/2)^2$. Par ailleurs, le produit $|R| \times |S|$ est d'autant plus grand que l'écart, en valeur absolue, de $|R|$ et de $|S|$ est petit.

Exemple : Nous allons déterminer le type de la partition de E, en *deux* classes, pour laquelle $|R \times S|$ est maximum.

Si m est le cardinal de l'une des classes, $(n-m)$ est le cardinal de l'autre classe. On supposera, sans restreindre en rien la généralité de la question, que m est plus grand que $(n-m)$: de sorte que le type de la partition aura la forme $(m, n-m)$, et que $|S| = m(n-m)$.

Nous allons déterminer la partition de l'entier n en deux nombres réels positifs ν et $(n-\nu)$,

$$(\nu > n - \nu) ,$$

pour lesquels $\nu(n-\nu) = n(n-1)/4$. Ces deux nombres sont les solutions de l'équation du second degré :

$$4x^2 - 4x\,n + n(n-1) = 0 ,$$

soit

$$\nu = (n + \sqrt{n})/2 \quad \text{et} \quad n - \nu = (n - \sqrt{n})/2 .$$

Si n est le carré d'un entier, $\sqrt{n}$ est un entier ; le type de la partition cherchée est alors exactement $((n+\sqrt{n})/2, (n-\sqrt{n})/2)$, n et $\sqrt{n}$ ayant même parité. Si n n'est pas le carré d'un entier, le type de la partition cherchée est

$$([(n+\sqrt{n})/2] + 1, [(n-\sqrt{n})/2]) ,$$

où $[x]$ désigne la partie entière de x.

Proposition 5 : si $k = 2$, la donnée de $|R(P)|$ est équivalente à celle du type (n_1, n_2) de la partition.

Comme d'habitude il suffit de raisonner sur Σn_i^2. Supposons donc qu'on ait $(n_1, n-n_1), (m_1, n-m_1)$ types de partitions différentes satisfaisant à

$$n_1^2 + (n-n_1)^2 = m_1^2 + (n-m_1)^2 .$$

On a $n_1 \geqslant n - n_1$, $m_1 \geqslant n - m_1$; de plus, si une des inégalités se réduit à une égalité, l'autre est

stricte, les types de partitions étant différentes. Donc $n_1 + m_1 > n$, et l'égalité ci-dessus s'écrit

$$(n_1 - m_1)(n_1 + m_1 - n) = 0$$

d'où $n_1 = m_1$, ce qui est absurde.

Ce dernier résultat nous amène à nous poser la question suivante : dans quelle mesure la donnée de $|R|$, soit de Σn_i^2 nous renseigne-t-elle sur le type de la partition ? D'une manière plus précise, étant donné un entier m positif, déterminer l'ensemble des types de partitions pour lesquels $\Sigma n_i^2 = m$, ainsi que son cardinal.

Proposition 6 : soit $\varphi(n, m, k)$ le cardinal de *l'ensemble des types* de partitions en k classes pour lesquelles $\Sigma n_i^2 = m$, alors on a

$$\varphi(n, m, k) = \sum_{0 \leqslant p \leqslant n/k} \varphi(n - p(k-1),\ m - 2pn + p^2(k-1),\ k-1).$$

Posons

$$\Phi(n, m, k) = \left\{ (n_1, \ldots, n_k) : n_1 \geqslant n_2 \geqslant \cdots \geqslant n_k,\ \sum_i n_i = n \quad \text{et} \quad \sum_i n_i^2 = m \right\},$$

donc

$$|\Phi(n, m, k)| = \varphi(n, m, k).$$

La formule s'obtient en partitionnant l'ensemble $\Phi(n, m, k)$ selon les valeurs de n_k, $1 \leqslant n_k \leqslant [n/k]$ (où $[x]$ est la partie entière de x). Un élément de $\Phi(n, m, k)$ pour lequel $n_k = h$ est caractérisé par :

$$n_1 \geqslant n_2 \geqslant \cdots \geqslant n_{k-1} \geqslant h, \qquad h + \sum_{i \leqslant k-1} n_i = n,$$

$$h^2 + \sum_{i \leqslant k-1} n_i^2 = m$$

ou encore par

$$n_1 - (h-1) \geqslant n_2 - (h-1) \geqslant \cdots \geqslant n_{k-1} - (h-1) \geqslant 1,$$

$$\sum_{i \leqslant k-1} (n_i - (h-1)) = n - (h-1)k - 1,$$

$$\sum_{i \leqslant k-1} (n_i - (h-1))^2 = m - 2(h-1)n + (h-1)^2 k - 1,$$

soit encore, en posant $\nu_i = n_i - (h-1)$, par

$$\nu_1 \geqslant \nu_2 \geqslant \cdots \geqslant \nu_{k-1} \geqslant 1,$$

$$\sum_{i \leqslant k-1} \nu_i = n - (h-1)k - 1,$$

$$\sum_{i \leqslant k-1} \nu_i^2 = m - 2(h-1)n + (h-1)^2 k - 1.$$

Le sous-ensemble de $\Phi(n, m, k)$ pour lequel $n_k = h$ peut donc être mis en correspondance avec

$$\Phi(n - (h-1)k - 1,\ m - 2(h-1)n + (h-1)^2 k - 1,\ k-1).$$

Il suffit de poser $p = k - 1$ pour établir la récurrence.

Pour déterminer les différents éléments de

$$\Phi(n, m, k),$$

on procède de proche en proche ; on commence par définir une partition de $\Phi(n, m, k)$, chaque partie étant définie par une valeur de n_k. Le sous-ensemble de $\Phi(n, m, k)$ pour lequel $n_k = h$ est déterminé par la donnée de l'ensemble

$$\Phi(n - (h-1)k - 1,\ m - 2(h-1)n + (h-1)^2 k - 1,\ k-1) \quad (1)$$

de la façon suivante ; à tout élément $(\nu_1, \nu_2, \ldots, \nu_{k-1})$ de (1) correspond, bijectivement, un élément

$$(n_1, n_2, \ldots, n_{k-1}, h)$$

de $\Phi(n, m, k)$ où $n_i = \nu_i + (h-1)$. De nouveau, pour déterminer l'ensemble (1), on commence par réaliser une partition selon la valeur prise par $\nu_{(k-1)}$, etc. Nous expliciterons la procédure sur un exemple que nous allons traiter ci-dessous. A elle seule, de la manière dont elle a été établie, l'équation de récurrence permet de déterminer $\Phi(n, m, k)$.

A partir de la donnée des ensembles $\Phi(n, m, k)$, on détermine l'ensemble des partitions de l'entier n, en un nombre de classes non fixé à priori, pour lesquelles $\sum_i n_i^2 = m$. En désignant cet ensemble par $\Psi(n, m)$, on a :

$$\Psi(n, m) = \sum_{k=1}^{n} \Phi(n, m, k) \quad \text{(somme d'ensembles)}$$

En désignant par $\psi(n, m)$ le cardinal de l'ensemble $\Psi(n, m)$, on a :

$$\psi(n, m) = \sum_{k=1}^{n} \varphi(n, m, k)$$

Pour que $\varphi(n, m, k)$ ne soit pas nul, il est nécessaire que m et n soient de la même parité.

En effet, $m + n$ se présente sous la forme :

$$m + n = \sum_{i=1}^{k} n_i^2 + \sum_{i=1}^{k} n_i = \sum_{i=1}^{k} n_i(n_i + 1) \quad ,$$

la dernière somme étant une somme finie d'entiers pairs, par ailleurs

$$r(q+1)^2 + (k-r)q^2 \leqslant m \leqslant (n-k+1)^2 + (k-1)$$

où q et r sont respectivement le quotient et le reste de la division de n par k ; $n = kq + r$ ($r < k$). Cette dernière relation résulte des propositions 2 et 3 ci-dessus.

Par ailleurs, notre problème est un aspect très particulier du problème de Warring. Dans le cadre de ce dernier problème, on a établi une formule asymptotique pour la quantité de représentations d'un nombre N comme somme de k puissances $s^{\text{ièmes}}$.

$$n_1^s + n_2^s + \cdots + n_k^s = N$$

Le cas qui nous intéresse correspond à $s = 2$

Exemple :

Nous nous proposons de déterminer l'ensemble $\Phi(20, 98, 5)$ dont le cardinal est $\varphi(20, 98, 5)$. Commencons par déterminer $\varphi(20, 98, 5)$ au moyen de la relation de récurrence. On a :

1/ $\varphi(20, 98, 5) = \varphi(19, 97, 4) + \varphi(14, 62, 4) + \varphi(9, 37, 4) + \varphi(4, 22, 4)$

a) $\varphi(19, 97, 4) = \varphi(18, 96, 3) + \varphi(14, 62, 3) + \varphi(10, 36, 3) + \varphi(6, 18, 3)$

a_1) $\varphi(18, 96, 3) = 0$ car $96 < 118 = 6^2 + 6^2 + 6^2$ [cf. ci-dessus]

a_2) $\varphi(14, 62, 3) = 0$ car $62 < 66 = 5^2 + 5^2 + 4^2$

a_3) $\varphi(10, 36, 3) = \varphi(9, 35, 2) + \varphi(6, 18, 2) + \varphi(3, 7, 2)$

a_{31}) $\varphi(9, 35, 2) = 0$ car $35 < 5^2 + 4^2 = 41$ [cf. ci-dessus]

a_{32}) $\varphi(6, 18, 2) = 1$ et la partition correspondante est (3,3), soit $\Phi(6, 18, 2) = \{(3,3)\}$

a_{33}) $\varphi(3, 7, 2) = 0$ car $7 > 2^2 + 1^2 = 5$ [cf. ci-dessus]

a_4) $\varphi(6, 18, 3) = \varphi(5, 17, 2) + \varphi(2, 8, 2)$

a_{41}) $\varphi(5, 17, 2) = 1$ et la partition correspondante est (4,1), soit, $\Phi(5, 17, 2) = \{(4,1)\}$

a_{42}) $\varphi(2, 8, 2) = 0$

b) $\varphi(14, 62, 4) = \varphi(13, 61, 3) + \varphi(9, 37, 3) + \varphi(5, 21, 3) + \varphi(1, 13, 3)$

b_1) $\varphi(13, 61, 3) = \varphi(12, 60, 2) + \varphi(9, 37, 2) + \varphi(6, 20, 2) + \varphi(3, 9, 2)$

b_{11}) $\varphi(12, 60, 2) = 0$ car $60 \leqslant 6^2 + 6^2 = 72$ [cf. ci-dessus]

b_{12}) $\varphi(9, 37, 2) = 0$ car $37 < 5^2 + 4^2 = 41$

b_{13}) $\varphi(6, 20, 2) = 1$ et la partition correspondante est (4,2), soit, $\Phi(6, 20, 2) = \{(4,2)\}$

b_{14}) $\varphi(9, 37, 3) = 0$ car $9 > 2^2 + 1^2 = 5$

b_2) $\varphi(9, 37, 3) = \varphi(8, 36, 2) + \varphi(5, 21, 2) + \varphi(2, 12, 2)$

Chacun des termes de la somme du second membre est nul, d'où :

$\varphi(9, 37, 3) = 0$

b_3) $\varphi(5, 21, 3) = 0$ car $21 > 3^2 + 1 + 1 = 11$

b_4) $\varphi(1, 13, 3) = 0$.

Comme il vient d'être fait, on peut vérifier que :

c) $\varphi(9,37,4) = 0$

d) $\varphi(4,22,4) = 0$

Nous allons maintenant déterminer les différents éléments de Φ

$$\Phi(20,98,5) = \left\{(n_1, n_2, n_3, n_4, n_5), \sum_1^5 n_i = 20 \ , \ \sum_1^5 n_i^2 = 98\right\}.$$

On a : $\Phi(6,18,2) = \{(3,3)\}$, d'où en reprenant les calculs à l'envers depuis a_{32}, on obtient :

$$\Phi(10,26,3) = \{(3+1, 3+1, 2)\} = \{(4,4,2)\},$$

$\varphi(6,18,2)$ étant le second terme dans le développement de $\varphi(10,36,3)$. De la même façon on a : $\Phi(6,18,3) = \{(4,1,1)\}$.

Il en résulte que

$$\Phi(19,97,4) = \{(4+2, 4+2, 2+2, 3), (4+3, 1+3, 1+3, 4)\}$$

puisque $\varphi(10,36,3)$ et $\varphi(6,18,3)$ sont respectivement le troisième et le quatrième terme dans le développement de $\varphi(19,97,4)$.

2/ $\Phi(19,97,4) = \{(6,6,4,3), (7,4,4,4)\}$

Par ailleurs, en considérant les calculs de b, on obtient, d'une manière analogue :

$$\Phi(13,61,3) = \{(4+2, 2+2, 3)\} = \{(6,4,3)\}.$$

D'où :

3/ $\Phi(14,62,4) = \{(6,4,3,1)\}$

Finalement en vertu de 1/, 2/ et 3/ on a :

$$\Phi(20,98,5) = \{(6,6,4,3,1), (7,4,4,4,1), (6+1, 4+1, 3+1, 1+1, 2)\}$$

soit

$$\Phi(20,98,5) = \{(6,6,4,3,1), (7,4,4,4,1), (7,5,4,2,2)\}.$$

IV — *ESPACES ULTRAMETRIQUES*

1 — Propriétés

Soit (E, d) un espace métrique i.e. un ensemble E et une fonction numérique positive

$$E \times E \longrightarrow R_+$$

vérifiant (M1) $d(x,y) = d(y,x)$

(M2) $d(x,y) = 0$ si et seulement si $x = y$

(M3) $d(x,z) \leqslant d(x,y) + d(y,z)$

On dira que c'est un espace ultramétrique si d vérifie en plus :

$$(UM)\quad d(x,z) \leqslant \max(d(x,y), d(y,z))$$

Il est trivial de vérifier que (UM) entraîne (M3) qui de ce fait devient superflue dans la définition d'un espace ultramétrique.

Rappelons quelques définitions :

— une *circonférence* de centre a et de rayon r dans un espace métrique E est l'ensemble

$$S(a;r) = \{x \in E : d(x,a) = r\}$$

(habituellement appelé "sphère").

— un *cercle circonférencié* $C(a;r)$ (resp. *non-circonférencié* $C(a;r^-)$) de centre a et de rayon r dans E (habituellement appelé boule fermée, resp. boule ouverte) est l'ensemble

$$C(a;r) = \{x \in E : d(x,a) \leqslant r\}$$

$$\text{resp. } C(a;r^-) = \{x \in E : d(x,a) < r\}$$

— un *diviseur* de E est une relation d'équivalence D dans E vérifiant

$$(DIV)\quad \forall a, b, x, y \in E \quad a\,D\,b \;\&\; (d(x,y) \leqslant d(a,b)) \longrightarrow x\,D\,y$$

Théorème 1 : un espace métrique (E, d) est ultramétrique si et seulement si tous les triangles sont isocèles, la longueur de la base étant inférieure ou égale à celle des côtés.

La propriété directe est évidente puisqu'elle implique (UM). Inversement soit x, y, z un triangle d'un espace ultramétrique (E, d) et soit x, y un des côtés les moins longs. On a :

$$d(x, z) \leqslant \max(d(x, y), d(y, z)) = d(y, z)$$

Permutons x et y

$$d(y, z) \leqslant d(x, z)$$

d'où

$$d(x, z) = d(y, z) \geqslant d(x, y)$$

Soit désormais (E, d) un espace ultramétrique.

Proposition 2 : tout point d'un cercle dans E, ultramétrique, en est centre.

Il suffit de considérer le triangle a, b, x (a centre du cercle donné, b élément de ce cercle, x élément du cercle de même rayon mais de centre b) qui est isocèle, et d'appliquer le thèorème 2.

Corollaire 2.1 : deux cercles non-disjoints sont concentriques.

Prendre un point quelconque de l'intersection comme centre de chacun.

Corollaire 2.2 : deux cercles de même rayon non-disjoints sont confondus.

Corollaire 2.3 : les cercles de même rayon forment une partition de l'espace ultramétrique E ; l'équivalence correspondante est un diviseur de E.

— une *valuation* d'un diviseur D de l'espace E est le nombre $v(D) = \sup_{xDy} d(x, y)$.

Il est facile de voir que si D et D' sont deux diviseurs de E, espace métrique fini, qui vérifient $D \leqslant D'$ dans le treillis $\mathcal{P}(E)$ alors $v(D) \geqslant v(D')$ et réciproquement.

Proposition 3 : C et C' étant des cercles disjoints dans E, la distance $d(x, y)$ d'un $x \in C$ et d'un $y \in C'$ ne dépend que de C et C' et non de x et y.

Considérer les triangles $x, \overline{x}, y$ et $\overline{x}, y, \overline{y}$ où $\overline{x} \in C'$ et appliquer le théorème 2.

Corollaire 3.1 : le quotient E/D d'un espace ultramétrique par un diviseur est un espace ultramétrique ; la distance de deux de ses points est strictement supérieure à $v(D)$ dans le cas fini.

Proposition 4 : tout cercle de rayon r^- dans E, non-disjoint avec une circonférence de rayon r dans E y est contenu

Considérer le triangle a, b, x où $b \in C \cap S$, $x \in C$, a centre de S et appliquer le théorème 2.

Corollaire 4.1 : une circonférence de rayon r dans un espace ultramétrique est la réunion des cercles (deux à deux disjoints) de rayon r^-.

— une *proximité ultramétrique* est une fonction numérique positive (éventuellement infinie)

$$p : E \times E \longrightarrow R_+ = R_+ \cup \{+\infty\} \text{ vérifiant}$$

(PUM1) $p(y, x) = p(x, y)$

(PUM2) $p(x, y) = +\infty$ si et seulement si $x = y$

(PUM3) $p(x, z) \geqslant \min(p(x, y), p(y, z))$.

Proposition 5 : si d est une distance ultramétrique, alors $-\log d$ est une proximité ultramétrique ; si p est une proximité ultramétrique, alors $\exp(-p)$ est une distance ultramétrique

— un *cercle de centre a et de petitesse* ω dans un espace ultramétrique E est l'ensemble

$$\{x \in E : p(x, a) \geqslant \omega\}$$

des points dont la proximité à a est supérieure à ω où ω est soit w soit w^- (dans ce cas l'inégalité est stricte)

— *l'ordre* d'un diviseur D de E est

$$w(D) = \inf_{xDy} p(x, y).$$

Il est trivial que $w(D) = -\log v(D)$.

2 — Chaines du treillis des partitions d'un ensemble fini et espaces ultramétriques associés

Soit E un ensemble fini. On va considérer des chaînes du treillis $\mathcal{P}(E)$, i.e. des suites finies $(\mathcal{P}_i)\ i \in I$ de partitions de E de moins en moins fines donc telles que pour tout $i \in I$, $P_i \leqslant P_{i+1}$ (I étant fini, on l'identifie à un segment initial de N). On ne modifie

en rien le problème en supposant que $P_0 = 0$, minorant universel de $\mathcal{R}$; c'est ce que nous ferons désormais.

Munissons l'ensemble $E \times E$ de la fonction d : $E \times E \longrightarrow I \subset N$ définie comme suit :

$d(x, y)$ est le plus petit $i \in I$ tel que x et y soient dans une même classe de P_i.

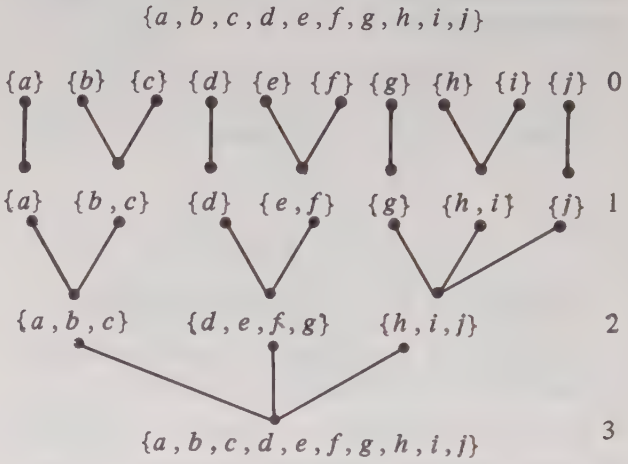
Proposition 1 : d est une distance ultramétrique sur E dont les diviseurs sont les P_i.

Les vérifications sont triviales.

Proposition 2 : inversement à tout espace ultramétrique fini (E, d) est associée canoniquement une chaîne de $\mathcal{R}$, i.e. une suite croissante de partitions de E, $(P_i)\, i \in I$ qui sont les diviseurs de E.

Par le Corollaire 3 de la Proposition 5, il suffit de prendre les cercles circonférenciés de rayon i et de faire parcourir N à i. En fait, E étant fini, i ne parcourra qu'un segment fini I de N, dont la longueur est inférieure ou égale à la hauteur de $\mathcal{R}$.

Exemple : considérons une chaîne du treillis de partitions défini sur l'ensemble



La matrice définissant la distance ultramétrique associée est la suivante :

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*
a	0									
b	2	0								
c	2	1	0							
d	3	3	3	0						
e	3	3	3	2	0					
f	3	3	3	2	1	0				
g	3	3	3	2	2	2	0			
h	3	3	3	3	3	3	3	0		
i	3	3	3	3	3	3	3	1	0	
j	3	3	3	3	3	3	3	2	2	0

La matrice étant symétrique, nous en avons représenté la moitié.

CHAPITRE PREMIER

MESURE DE SIMILARITE ET PRÉORDONNANCE

I — *INTRODUCTION*

1 — Caractères et attributs

Nées à propos de la taxinomie animale, les méthodes de la classification automatique s'appliquent avec fécondité aux domaines aussi multiples et variés que sont : la sociologie végétale, l'archéologie, la géographie, la sociologie, la psychologie, l'économie, etc. C'est relativement à une visée classificatoire qu'on cherchera à découvrir une partition ou une chaîne de partitions qui "respecte d'une manière satisfaisante les ressemblances entre objets" sur une population finie, E, d'objets, donnée. Nous attendrons le chapitre suivant pour préciser certains aspects de ce problème.

On suppose établi un ensemble fini de caractères $\{a_1, a_2, \ldots, a_h, \ldots, a_p\}$, répondant à un but taxinomique et suffisant pour décrire les objets de E. L'ensemble des différentes modalités (ou valeurs) d'un caractère a_h sera désigné par A_h. $a_1, \ldots, a_h, \ldots, a_p$ se présentent sous la forme des différentes composantes de l'application a qui définit la description de E dans le produit cartésien

$$A_1 \times A_2 \times \ldots \times A_h \times \ldots \times A_p \,.$$

$$E \xrightarrow{a} A_1 \times A_2 \times \ldots \times A_h \times \ldots \times A_p$$

$$x \in E \longrightarrow a(x) = (a_1(x), a_2(x), \ldots, a_h(x), \ldots, a_p(x))$$

où $a_h(x)$ est la modalité du caractère a_h présente chez x.

Nous nous attacherons essentiellement au cas où chacun des ensembles A_h est fini et sans structure. C'est là une situation très générale dans les Sciences humaines. Certains des ensembles A_h peuvent, naturellement, être munis de structures plus ou moins riches ; A_h peut par exemple être un ensemble ordonné, un espace métrique ou un espace vectoriel muni d'un produit scalaire. Une prochaine publication de N. Jardine et R. Sibson traitera précisément de l'établissement d'une mesure de similarité dans le cas où les divers ensembles A_h sont respectivement munis de diverses structures. Toutefois, les méthodes que nous étudierons dans le cas général, défini ci-dessus, pourront aisément être adaptées à des structures plus riches.

Soient $a_h^1, a_h^2, \ldots, a_h^i, \ldots, a_h^k$ les différentes modalités du caractère a_h. On peut, relativement à la description des objets de E, s'intéresser à certaines seulement d'entre elles, pour des raisons de pertinence ou techniques. L'attribution d'une modalité a_h^i à un objet donné x sera représentée au moyen d'une variable logique qui vaut 1 si x possède a_h^i et 0 sinon.

1.1 — *Définition.* Nous appellerons attribut toute modalité d'un caractère retenue pour la description de E.

Soit $\{\alpha_1, \alpha_2, \ldots, \alpha_T\} = A$ l'ensemble des attributs ainsi établi. Un objet x est alors défini par la donnée du sous-ensemble X, $(X \subset A)$ des attributs qu'il possède ; il est représenté dans $\{0, 1\}^T$ par le point $\alpha(x) = (x_1, x_2, \ldots, x_t, \ldots, x_T)$ où x_t est une variable logique qui vaut 1 si l'objet x possède l'attribut α_t et 0 sinon.

Ces données sont en général consignées en un tableau qui est une matrice d'incidence, où les lignes représentent les attributs et les colonnes les objets ; les différentes colonnes sont respectivement définies par les différents points $\alpha(x)$, $x \in E$.

2 — Signification d'un attribut pour une classification

Soit $(E_1, E_2, \ldots, E_j, \ldots, E_k)$ une partition de la population d'objets E qui répond à un but classificatoire donné. Si α est un attribut donné, désignons par f_j la proportion des objets de la classe E_j qui possèdent l'attribut α. $(f_1, f_2, \ldots, f_k)$ définira la distribution de la fréquence relative de présence de α sur les différentes classes.

L'attribut α explique (ou discrimine) d'autant mieux la classification que la dispersion de $j \rightsquigarrow f_j$ est plus grande par rapport à la dispersion de la fréquence de présence de α dans E.

2.1 — *Définition*

Nous appellerons *signification* de l'attribut α par rapport à la classification $(E_1, E_2, \ldots, E_j, \ldots, E_k)$ la quantité

$$\frac{1}{k-1} \sum_{j=1}^{k} (f_j - \bar{f})^2 \Big/ \frac{\bar{f}(1-\bar{f})}{n} \qquad (1)$$

où $\bar{f} = \frac{1}{k} \sum_{j=1}^{k} f_j$; rapport de l'estimation de la variance interclasse sur celle, globale. Dans la mesure où les cardinaux des différentes classes E_j sont assez grands, $(n_j \bar{f}(1-\bar{f}) \geqslant 15)$; et, dans l'hypothèse de l'absence de signification de l'attribut α, la statistique (1) peut être considérée comme une réalisation d'une variable aléatoire de Fisher-Snédécor avec $(k-1)$ et $(n-1)$ degrés de liberté.

Il en résulte une pondération des attributs où à chaque α_i, $i = 1, 2, \ldots, T$, est attaché un coefficient positif tel que (1).

Il est intéressant de déterminer ceux des attributs dont la variance interclasse est significativement grande.

Un attribut α définit sur E une partition en deux classes : celle dont les éléments possèdent α et celle complémentaire; S. Régnier préfère définir une autre pondération qui à α affecte le χ^2 calculé sur le tableau qui croise la partition $(E_1, E_2, \ldots, E_k)$ et celle définie par l'attribut.

Exemple : Une classification obtenue par diagnostic sépare 178 malades mentaux en 3 classes : N, P et S ;

N : névrose

P : psychose non schizophrénique

S : psychose schizophrénique

On étudie la présence de la couleur noire dans des tableaux non figuratifs peints par les malades. Les résultats sont les suivants :

Diagnostics

		N	P	S	Total
Couleur noire	présente	*37*	10	48	95
	absente	14	*17*	52	83
	Total	51	27	100	178

Le χ^2 calculé ($\chi^2 = 11{,}59$) est significatif. La couleur noire est donc plus employée par les névrosés, moins par les psychoses non schizophréniques. Ce résultat est également obtenu par le calcul de la statistique (1) pour laquelle on obtient : 8,02 ; valeur hautement significative en regard de la distribution de (1) dans l'hypothèse d'absence de discrimination de l'attribut.

La donnée d'une chaîne de partitions définissant une hiérarchie de classifications permet pour chaque attribut de déterminer la distribution de la statistique (1) le long de la chaîne. Cette distribution indique le niveau de la hiérarchie que l'attribut discrimine le mieux.

3 — Choix des attributs

Une question particulièrement délicate concerne le choix des attributs pertinents vis-à-vis du problème étudié et l'importance qu'il convient d'accorder à chacun d'entre eux pour définir au mieux la classifi-

cation ou la hiérarchie de classifications. Il semble qu'il y ait là un cercle vicieux, la classification que l'on voudrait obtenir est précisément celle qu'il faudrait connaître à l'avance pour définir la bonne pondération [cf. § 2 ci-dessus].

Plaçons-nous, par exemple, dans le cadre de la taxinomie animale. Une description exhaustive des espèces est impossible ; toutefois, un échantillonnage plus ou moins arbitraire des attributs, permet, selon Sokal et Sneath [cf. 32] d'aboutir à une classification, rendant possible la connaissance des espèces, en vertu de deux hypothèses :

a) L'hypothèse du nexus, selon laquelle tout attribut est affecté par plusieurs facteurs génétiques et, réciproquement, la plupart des gènes affectent plusieurs caractères.

De la sorte tout attribut devrait nous informer à propos de plusieurs gènes et il est possible de retrouver les effets d'un gène à travers plusieurs attributs.

b) L'hypothèse de la non-spécifité, selon laquelle il n'existe pas de groupe de gènes dont l'influence se limiterait exclusivement à une seule partie, ou une seule fonction de l'organisme.

Ainsi, si un attribut est affecté par plusieurs facteurs génétiques, sa présence et celle de beaucoup d'autres sont simultanées et on peut espérer que cela apparaisse dans l'étude des ressemblances.

Des correspondants des hypothèses (a) et (b), qui postulent l'existence d'une famille finie de facteurs, semblent se retrouver dans de nombreux domaines des Sciences humaines.

Finalement, puisque les attributs ne s'imposent pas d'eux-mêmes, que l'importance relative d'un attribut est inconnue à priori, que dans la plus vaste classification que l'homme ait édifiée, celle du règne animal, l'étude des espèces se déchiffre dans leurs ressemblances, l'étude des effets, des ressemblances est donc ce qu'il y a lieu d'entreprendre, une fois que le spécialiste aura déterminé les attributs de description.

Pour la plupart des méthodes de classification, l'information dont on dispose est relative à la ressemblance des objets pris deux par deux. Cette information peut se présenter sous la forme d'une mesure de la ressemblance définie au moyen d'une fonction réelle positive sur $E \times E$; elle peut être simplement la donnée pour tout couple de paires, (p, q), de celle des deux paires p ou q formée des objets les plus ressemblant ; en d'autres termes cette information est un préordre total, ω, sur l'ensemble F des paires d'objets distincts de E, appelé préordonnance sur E. Pour un tel préordre une paire p précède une paire q $(p \leqslant q)$ si les objets composant la paire p se ressemblent davantage que ceux composant q. La préordonnance est généralement associée au choix d'une mesure de similarité. [cf. III].

Nous allons maintenant donner une définition axiomatique d'une mesure de similarité que nous comparerons avec une autre définition plus générale. Notre définition nous permettra d'étudier la variabilité de la préordonnance associée quand on remplace une similarité par une autre. Nous terminerons en proposant certaines préordonnances particulières.

II — *DEFINITION D'UNE MESURE DE SIMILARITE*

1 — Généralités

Soient x et y deux objets quelconques de E, $\alpha(x) = (x_1, x_2, \ldots, x_t, \ldots, x_T)$ et

$$\alpha(y) = (y_1, y_2, \ldots, y_t, \ldots, y_T)$$

les deux points les représentant dans $\{0, 1\}^T$.

Relativement à un même attribut α_t, les deux objets x et y sont dits avoir une association *positive* (resp. négative), si α_t est présent (resp. absent) simultanément chez les deux objets ; c'est-à-dire $x_t y_t = 1$ (resp. 0).

Compte tenu de ce que nous avons exprimé, au paragraphe I.3, à la suite des hypothèses (a) et (b) ; nous résignant, dans une première approche, à l'ignorance, de la manière la plus objective, nous supposerons les conditions suivantes.

1.1 — *Conditions générales*

La contribution d'une association positive (resp. négative) à la mesure de la ressemblance de deux objets est la même pour tout attribut. De plus cette contribution pour une association négative est au plus égale à celle pour une association positive.

Aux objets x et y associons les cardinaux suivants :

s (resp. t) cardinal du sous-ensemble des attributs possédés en commun (resp. non possédés par aucun des deux objets).

s est le nombre d'associations positives et t, négatives.

u (resp. v) cardinal du sous-ensemble des attributs possédés par l'objet x (resp. y) et non possédés par y (resp. x).

Désignons par X et Y les sous-ensembles d'attributs possédés respectivement par x et y. On a

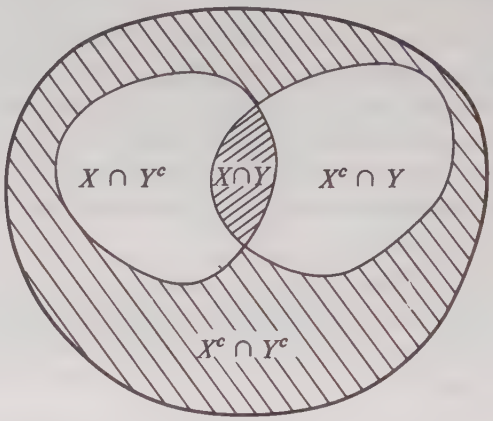
$$s = |X \cap Y| = \sum_{i=1}^{T} x_i y_i \quad , \quad t = |X^c \cap Y^c| = \sum_{i=1}^{T} (1 - x_i)(1 - y_i)$$

X^c (resp. Y^c) étant le complémentaire dans A de X (resp. Y)

$$u = |X \cap Y^c| = \sum_{i=1}^{T} x_i(1 - y_i) \quad , \quad v = |X^c \cap Y| = \sum_{i=1}^{T} (1 - x_i)y_i$$

On a : $s + u = |X|$ et $s + v = |Y|$.

De plus s, u, v et t sont liés par la relation $s + u + v + t = T$.



2 — Définitions et Propriétés

2.1 — *Définition.* Nous appelons indicateur du couple (x, y), $I(x, y)$, le triplet (s, u, v) où s, u et v sont relatifs au couple (x, y).

Nous considérons $I(x, y) = (s, u, v)$, comme la donnée de base de la mesure de la ressemblance des deux objets.

Exemple : soit le tableau des données où

$$E = \{a, b, c, d, e, f\} ;$$

	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9	α_{10}	α_{11}	α_{12}	α_{13}	α_{14}	α_{15}	α_{16}	α_{17}	α_{18}
a	0	1	0	1	1	0	0	1	0	1	1	0	1	0	0	1	1	0
b	0	0	0	0	0	0	1	0	0	1	0	0	0	1	1	0	1	1
c	0	0	1	0	0	0	1	0	0	1	1	1	1	0	1	0	0	1
d	0	1	0	1	0	1	0	0	0	0	1	0	1	1	0	0	1	0
e	1	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0
f	1	0	0	1	1	1	0	1	1	0	1	0	0	1	0	1	1	0

et le tableau donnant pour chaque couple d'objets son indicateur (s, u, v) ainsi que la valeur de $u + v$ correspondante.

On remplira les cases vides du tableau en tenant compte de la relation : $I(x, y) = (s, u, v)$ si et seulement si $I(y, x) = (s, v, u)$.

	a	*b*	*c*	*d*	*e*	*f*
a	(9,0,0) 0	(2,7,4) 11	(3,6,5) 11	(5,4,2) 6	(1,8,4) 12	(6,3,4) 7
b		(6,0,0) 0	(4,2,4) 6	(2,4,5) 9	(2,4,3) 7	(2,4,8) 12
c			(8,0,0) 0	(2,6,5) 11	(2,6,3) 9	(1,7,9) 16
d				(7,0,0) 0	(2,5,3) 8	(5,2,5) 7
e					(5,0,0) 0	(3,2,7) 9
f						(10,0,0) 0

2.2 — *Définition 1.* Une mesure de similarité est une fonction réelle positive S définie sur l'ensemble $E \times E$ qui se présente sous la forme :

$$(x, y) \longrightarrow S(x, y) = \mathfrak{S}(I(x, y)) = \mathfrak{S}(s, u, v)$$

où $\mathfrak{S}(s, u, v)$ est une fonction définie sur le sous-ensemble de N^3 : $\{(s, u, v) ; s + u + v \leqslant T\}$, croissante par rapport à s, symétrique en u et v et décroissante par rapport à u, la croissance par rapport à s ou la décroissance par rapport à u étant stricte.

Nous allons comparer notre définition avec une définition plus générale due à W.F. de la Véga. (cf. [34]).

2.3 — *Définition 2.* Une mesure de similarité est une fonction réelle positive S définie sur l'ensemble $E \times E$ qui satisfait aux conditions

1/ $\forall (x, y) \in E \times E : S(x, y) = S(y, x)$

2/ Si x et y diffèrent seulement du point de vue de a_k, ($x_i = y_i$ pour tout $i \neq k$ et $x_k \neq y_k$), et si z est tel que $|z_k - x_k| < |z_k - y_k|$; alors

$$S(x, z) \geqslant S(y, z) .$$

2.4 — *Proposition.* Toute similarité satisfaisant aux axiomes de la définition 1, satisfait aux axiomes de la définition 2. Réciproquement, toute similarité, satisfaisant aux axiomes de la définition 2, qui se met sous la forme $S(x, y) = \mathfrak{S}(s, u, v)$, satisfait à ceux de la définition 1.

Il est aisé de démontrer la première partie de la proposition. Occupons-nous de la réciproque. Si un indice de similarité, $S(x, y)$, se met sous la forme : $S(x, y) = \mathfrak{S}(I(x, y)) = \mathfrak{S}(s, u, v)$, l'axiome 1 de la "définition 2" implique la symétrie de la fonction $\mathfrak{S}(s, u, v)$ en u et v : $\mathfrak{S}(s, u, v) = \mathfrak{S}(s, v, u)$. D'autre part l'axiome 3 de la "définition 2" s'exprime par l'inégalité

$$\mathfrak{S}(s, u, v) \geqslant \mathfrak{S}(s - \epsilon^2 , u + (1 - \epsilon)^2 , v + \epsilon^2) \quad (1)$$

où $\epsilon = 0$ ou 1.

En faisant $\epsilon = 0$, on obtient la condition de décroissance par rapport à u de la fonction $\mathfrak{S}(s, u, v)$.

En faisant $\epsilon = 1$, on obtient :

$$\mathfrak{S}(s, u, v) \geqslant \mathfrak{S}(s - 1, u, v + 1) . \quad (2)$$

On a nécessairement

$$\mathfrak{S}(s, u, v) \geqslant \mathfrak{S}(s - 1, u, v) . \quad (3)$$

Car sinon on aurait

$$\mathfrak{S}(s - 1, u, v) > \mathfrak{S}(s, u, v) ,$$

ce qui implique en raison de (2)

$$\mathfrak{S}(s - 1, u, v) > \mathfrak{S}(s - 1, u, v +) ,$$

ce qui est impossible en vertu de la décroissance de $\mathfrak{S}(s, u, v)$ par rapport à u et de la symétrie de $\mathfrak{S}(s, u, v)$ en u et v, déjà établies ci-dessus. La condition (3) exprime que la fonction $\mathfrak{S}(s, u, v)$ est croissante par rapport à s. C.Q.F.D.

Notre définition est donc plus restrictive, la restriction consiste exactement à imposer à l'indice de similarité d'être une fonction du triplet (s, u, v). Notre définition reste assez générale puisqu'elle inclut les différents indices de similarité, proposés, qui sont, à une élévation au carré près, des fractions rationnelles en s, $u + v$ et uv. [cf. ci-dessous].

2.5 — *Expression de différentes mesures de similarité proposées par différents auteurs*

(1) RUSSEL et RAO	(1940)	$S(x, y) = \frac{s}{T}$
(2) M.G. KENDALL ; SOKAL MICHENER	(1958)	$S(x, y) = 1 - \frac{u + v}{T}$
(3) ROGER et TANIMOTO	(1960)	$S(x, y) = \frac{T - (u + v)}{T + (u + v)}$
(4) HARMAN	(1961)	$S(x, y) = 1 - \frac{2(u + v)}{T}$
(5) JACCARD	(1908)	$S(x, y) = \frac{s}{s + (u + v)}$
(6) KULEZYNSKI	(1927)	$S(x, y) = \frac{s}{u + v}$
(7) DICE	(1945)	$S(x, y) = \frac{s}{s + \frac{1}{2}(u + v)}$
(8) SOKAL et SNEATH		$S(x, y) = \frac{s}{s + 2(u + v)}$
(9) KULEZYNSKI		$S(x, y) = \frac{1}{2}\left[\frac{s}{s + u} + \frac{s}{s + v}\right]$
(10) OCHIAI	(1957)	$S(x, y) = \frac{s}{\sqrt{(s + u)(s + v)}}$
(11) YULE	(1911)	$S(x, y) = \frac{st - uv}{st + uv}$
(12) PEARSON		$S(x, y) = \frac{st - uv}{\sqrt{(s + u)(s + v)(t + u)(t + v)}}$

où $I(x, y) = (s, u, v)$.

Notons que la similarité (1) est une fonction de s seulement, que les indices (2), (3) et (4) sont respectivement des fonctions de $d = u + v$ et que (5), (6), (7) et (8) sont des similarités dont chacune est une fonction du rapport $s/(u + v)$. Remarquons encore que la similarité (9) est la moyenne algébrique de $s/(s + u)$ et de $s/(s + v)$ et que 10 en est la moyenne géométrique.

Notons enfin que $S(x, x) = \mathfrak{S}(s, 0, 0) = 1$ pour ces différentes mesures de similarité proposées à l'exeption de (1) et de (6) où

$$S(x, x) = \frac{1}{T}\sum_{i=1}^{T} x_i^2$$

pour (1) et $S(x, x) = \infty$ pour (6).

Relativement aux indices (1) et (2), $1 - \frac{s}{T}$ et $\frac{u + v}{T}$ définissent deux distances sur l'ensemble des parties de A.

Ces similarités ont été proposées dans le cas où les différents caractères sont bivalents et où l'ensemble des attributs, A, est déterminé en retenant pour chacun des caractères celle des deux modalités la plus significatives.

III — *PREORDONNANCE ASSOCIEE A UNE MESURE DE SIMILARITE*

1 — Définitions et propriétés générales

Le choix d'une mesure de similarité S définit, sur l'ensemble des paires d'objets distincts de E, F, un préordre total :

$(\{x, y\}, \{z, t\}) \in F \times F$:

$$\{x, y\} \leqslant \{z, t\} \Longleftrightarrow S(x, y) \geqslant S(z, t) \ .$$

Ce préordre est la "préordonnance sur E associée à S", nous la noterons $\omega(S)$; d'autre part, nous noterons encore $\omega(S)$ le graphe dans $F \times F$ de ce préordre. Comme nous l'avons annoncé, *la donnée de base de certaines méthodes de classification est précisément* $\omega(S)$.

Sous ce point de vue, qui est le nôtre, on saisit l'arbitraire qui consiste à prendre pour mesure de similarité, précisément une certaine *fraction rationnelle* en s, $u + v$ et uv. Devant l'incertitude où on se trouve, lorsqu'on a à choisir l'une ou l'autre des similarités, on peut se demander *dans quelle mesure la préordonnance* $\omega(S)$ *varie lorsqu'on remplace une mesure par une autre.*

Exemple : dans le cas du tableau précédent [cf. § II.2], on a pour ω_s :

$$\begin{aligned}\{a, f\} < \{a, d\} = \{d, f\} < \{b, c\} < \{a, c\} = \{e, f\} \\ < \{a, b\} = \{b, d\} < \{c, d\} = \{b, e\} = \{c, e\} \\ = \{d, e\} = \{b, f\} < \{c, f\} = \{a, e\},\end{aligned}$$

les valeurs correspondantes de s étant respectivement : 6, 5, 4, 3, 2 et 1 et pour ω_{s+t} :

$$\begin{aligned}\{a, d\} = \{b, c\} < \{b, e\} = \{a, f\} = \{d, f\} < \{d, e\} \\ < \{b, d\} = \{c, e\} = \{e, f\} < \{a, b\} = \{a, c\} \\ = \{c, d\} < \{a, e\} = \{b, f\} < \{c, f\},\end{aligned}$$

les valeurs de $(s + t)$ correspondantes étant respectivement : 12, 11, 10, 9, 7, 6 et 2 ; donc celles de $(u + v)$: 6, 7, 8, 9, 11, 12 et 16.

1/ *Définition.* Deux mesures de similarité sont équivalentes sur un ensemble E d'objets donné, si et seulement si, les préordonnances respectivement associées sur E sont identiques.

$S(x, y)$ et $S'(x, y)$ désignant deux similarités, il y a lieu de remarquer que s'il existe une fonction numérique strictement croissante, f, telle que

$$S'(x, y) = f[S(x, y)] \ ,$$

S et S' sont équivalents sur tout ensemble, E, d'objets.

2/ *Proposition.* Si le nombre d'attributs possédés par un même objet de E est invariable, toutes les similarités sont équivalentes sur E.

La fonction, $\mathcal{S}(s, u, v)$ où $S(x, y) = \mathcal{S}(s, u, v)$ est supposée strictement croissante en s, ou, non exclusivement, strictement décroissante en u.

Dans les conditions de la proposition, l'indicateur du couple (x, y) de $E \times E$ se présente sous la forme : $I(x, y) = (s, l - s, l - s)$ où $s = \sum_{i=1}^{T} x_i y_i$ et l est le nombre d'attributs commun à tous les éléments de E. Il en résulte qu'une mesure de similarité quelconque, $S(x, y) = \mathcal{S}(s, u, v)$, se met sous la forme

$$\mathcal{S}(s, l - s, l - s) = R(s) \ ;$$

où $R(s)$ est une fonction numérique à valeurs positives, définie sur $\{0, 1, 2, \ldots, T\}$ et strictement croissante.

2 — Ecart entre deux préordonnances associées à deux similarités

En vue de l'étude de la stabilité, de la préordonnance associée, dans le choix d'une similarité, lorsque deux indices S et S' ne sont pas équivalents, il y a lieu de définir l'"écart" entre les deux préordonnances, respectivement associées, $\omega(S)$ et $\omega(S')$, sur E.

2.1 — *Définitions*

a) ω désignant le graphe dans $F \times F$ d'un préordre total sur F, introduisons relativement au graphe ω

$$\omega_1 = \{(p\,,q) \in F \times F : ((p\,,q) \in \omega \quad \text{et} \quad \neg((q\,,p) \in \omega)\}$$

et

$$\omega_2 = \{(p\,,q) : (q\,,p) \in \omega_1\}$$

(on rappelle que le signe $\neg$, placé devant un énoncé, le change en sa négation). Si ω est relatif à un ordre total, ω_1 coïncide avec ω et ω_2 avec le complémentaire de ω dans $F \times F$.

b) ω et ω' désignant les graphes dans $F \times F$ de deux préordres totaux, nous appelons "désaccord" entre ω et ω', le sous-ensemble de $F \times F$.

$$\Delta(\omega\,,\omega') = \omega_1 \cap \omega'_2 + \omega_2 \cap \omega'_1$$

où ω'_1 et ω'_2 sont associés à ω' de la même façon que le sont ω_1 et ω_2 à ω.

Remarquons que :

$$(p\,,q) \in \Delta(\omega\,,\omega') \Longleftrightarrow (q\,,p) \in \Delta(\omega\,,\omega')\,.$$

c) On définira l'"écart" entre ω et ω' par le nombre entier $\frac{1}{2}|\Delta(\omega\,,\omega')|$; $|\Delta(\omega\,,\omega')|$ est un entier pair en vertu de la remarque faite. $\Delta(\omega\,,\omega')$ est l'ensemble des couples $(p\,,q)$ pour lesquels p et q ne sont pas comparables pour l'intersection des deux préordres. Dans le cas où ω et ω' sont des ordres totaux, $\Delta(\omega\,,\omega')$ est la différence symétrique entre ω et ω'.

L'écart entre ω et ω' est *le nombre d'inversions* que présente ω' par rapport à ω.

Lorsque $\omega = \omega(S)$ et $\omega' = \omega(S')$, le couple $(\{x\,,y\}\,,\{z\,,t\})$ donne lieu à une inversion si et seulement si $S(x\,,y) < S(z\,,t)$ et $S'(x\,,y) > S'(z\,,t)$ ou $S(x\,,y) > S(z\,,t)$ et $S'(x\,,y) < S'(z\,,t)$.

d) Nous désignerons dans la suite par $\underline{\omega}$ et $\overline{\omega}$ les préordonnances sur E, respectivement associées aux mesures de similarité $S(x\,,y) = s$ et $S'(x\,,y) = s+t$ et par Δ le "désaccord" entre $\underline{\omega}$ et $\overline{\omega}$; $\Delta = (\underline{\omega}\,,\overline{\omega})$.

Reprenons l'exemple précédent :

$$\Delta = \{(\{a\,,f\}\,,\{a\,,d\})\,,(\{a\,,d\}\,,\{a\,,f\})\,,$$
$$(\{a\,,c\}\,,\{b\,,d\})\,,(\{b\,,d\}\,,\{a\,,c\})\,,(\{a\,,c\}\,,\{b\,,e\})\,,$$
$$(\{b\,,e\}\,,\{a\,,c\})\,,(\{a\,,c\}\,,\{c\,,e\})\,,(\{c\,,e\}\,,\{a\,,c\})\,,$$
$$(\{a\,,c\}\,,\{d\,,e\})\,,(\{d\,,e\}\,,\{a\,,c\})\,,(\{e\,,f\}\,,\{b\,,e\})\,,$$
$$(\{b\,,e\}\,,\{e\,,f\})\,,(\{e\,,f\}\,,\{d\,,e\})\,,(\{d\,,e\}\,,\{e\,,f\})\,.$$

Nous allons désormais restreindre notre attention à l'ensemble des mesures de similarité qui se présentent sous la forme assez générale $\mathfrak{S}(s\,,u+v)$.

2.2 — *Proposition 1.* α et β étant les préordonnances respectivement associées aux deux similarités

$$\varphi(s\,,u+v) \quad \text{et} \quad \psi(s\,,u+v)\,,$$

on a : $\Delta(\alpha\,,\beta) \subset \Delta$.

En effet, soit $(\{x\,,y\}\,,\{z\,,t\})$ un élément de $\Delta(\alpha\,,\beta)$, désignons par

$$(s\,,u\,,v) \quad \text{et} \quad (s+r\,,u+k\,,v+h)\,,$$

les indicateurs des couples $(x\,,y)$ et $(z\,,t)$. Dans ces conditions,

$$\varphi(s+r\,,u+v+k+h) - \varphi(s\,,u+v)$$

et

$$\psi(s+r\,,u+v+k+h) - \psi(s\,,u+v)$$

sont non nuls et de signes contraires. Il en résulte que r et $(k+h)$ sont non nuls et de même signe, donc $(\{x\,,y\}\,,\{z\,,t\})$ appartient à Δ.

2.3 — *Remarque importante*

Il résulte de cette proposition que, parmi les indices de similarité qui se présentent sous la forme $\mathfrak{S}(s\,,u+v)$, les deux indices pour lesquels les préordonnances respectivement associées ont l'"écart" le plus grand sont $S(x\,,y) = s$ et $S'(x\,,y) = s+t$.

2.4 — *Proposition 2.* Si le nombre d'attributs possédés par un même objet de E prend l'une de deux valeurs consécutives (l ou $l+1$), l'ensemble Δ est vide : $\Delta = \emptyset$.

En effet si Δ n'est pas vide, il existerait un élément de $F \times F$, $(\{x\,,y\}\,,\{z\,,t\})$ pour lequel :

$$\sum_i x_i y_i > \sum_i z_i t_i \quad \text{et} \quad \sum_i (x_i + y_i) > \sum_i (z_i + t_i) + 2\left(\sum_i x_i y_i - \sum_i z_i t_i\right) \quad (1)$$

La relation (1) implique

$$\sum_i (x_i + y_i) - \sum_i (z_i + t_i) \geqslant 3\,. \qquad (2)$$

Or $\sum_i (x_i + y_i) \leqslant 2l+2$ et $\sum_i (z_i + t_i) \geqslant 2l$, ce qui rend (2) impossible.

Remarque. Dans les conditions de la proposition précédente, le nombre d'inversions que présente $\underline{\omega}$ par rapport à $\overline{\omega}$ est nul.

2.5 — *Proposition 3.* V désignant la variance dans E du nombre d'attributs possédés par un même objet, on a :

$$\frac{|\Delta|}{|F \times F|} < \frac{4V}{9}$$

Si l_x désigne le nombre d'attributs possédés par l'objet x , $x \in E$, on a

$$V = \frac{1}{n} \sum_{x \in E} (l_x - \bar{l})^2 \quad \text{où} \quad \bar{l} = \frac{1}{n} \sum_{x \in E} l_x$$

L'énoncé de la proposition exprime *que l'écart entre* $\underline{\omega}$ *et* $\overline{\omega}$ *est strictement majoré* par $|F|^2 \times \frac{2V}{9}$.

Désignons par $\Phi(p, q)$, $(p, q) \in F \times F$, la fonction indicatrice de l'ensemble Δ. On a :

$$|\Delta| = \sum_{G} \Phi(\{x, y\}, \{z, t\}) + \sum_{H} \Phi(\{x, y\}, \{x, y\}) . \tag{1}$$

La première somme, (de gauche à droite), est étendue à l'ensemble G, $G \subset F \times F$, des couples de paires sans composante commune, et la seconde somme à l'ensemble H ; $H \subset F \times F$, des couples de paires ayant une composante commune. On a pour les cardinaux de G et de H :

$$|G| = n(n-1)(n-2)(n-3)/4 \tag{2}$$
$$|H| = n(n-1)(n-2) .$$

Posons :

$$p = \frac{|G \cap \Delta|}{|G|} \quad \text{et} \quad p' = \frac{|H \cap \Delta|}{|H|} .$$

Dans ces conditions la relation (1) peut être écrite sous la forme

$$|\Delta| = \frac{1}{4} n(n-1)(n-2)(n-3)\, p + n(n-1)(n-2)\, p' . \tag{3}$$

Munissons l'ensemble E de la loi uniforme, le nombre l_x d'attributs possédés par un objet x pris dans E selon le modèle, (équiprobabilité d'apparition de chacun des objets), est une variable aléatoire de moyenne $\bar{l}$ et de variance V. Dans ce cadre, considérons un échantillon de quatre objets indépendants, x , y , z et t de E ; p est précisément la probabilité de l'événement : $(\{x, y\}, \{z, t\}) \in \Delta$. Un tel événement implique :

$$\{|(l_x + l_y) - (l_z + l_t)| \geqslant 3\} .$$

Donc :

$$p \leqslant \Pr\{|(l_x + l_y) - (l_z + l_t)| \geqslant 3\} .$$

Or en vertu de l'inégalité de Tchebychef

$$\Pr\{|l_x + l_y) - (l_z + l_t)| \geqslant 3\} \leqslant \frac{1}{3^2} E[(l_x + l_y) - (l_z + l_t)]^2 = \frac{4V}{9}$$

D'autre part, pour un échantillon de 3 objets indépendants, x, y et z de E, p' est précisément la probabilité de l'événement : $(\{x, y\}, \{x, z\}) \in \Delta$. Un tel événement implique :

$$\{|l_y - l_z| \geqslant 3\} .$$

Donc :

$$p' \leqslant \Pr\{|l_y - l_z| \geqslant 3\} \leqslant \frac{1}{3^2} E(l_y - l_z)^2 = \frac{2V}{9} .$$

Il en résulte que $|\Delta|$, défini en (3), est majoré par

$$\frac{1}{9} n(n-1)(n-2)(n-3)\, V + \frac{2}{9} n(n-1)(n-2)\, V$$

soit

$$\frac{1}{9} (n-2)(n-1)^2\, nV . \tag{4}$$

Le cardinal de $F \times F$ étant $\frac{1}{4}(n-1)^2 n^2$, on a

$$\frac{|\Delta|}{F \times F} \leqslant \frac{4}{9} \times \frac{n-2}{n} V < \frac{4V}{9} . \tag{5}$$

Ce dernier résultat, comme celui de Tchebychef qui a servi à l'établir, a un intérêt seulement théorique ; il exprime que si la variance dans E du nombre

d'attributs possédés par un même objet est petite, le nombre d'inversions entre les préordonnances associées à deux similarités quelconques est petit. Dans la pratique, il arrive souvent (questionnaires, codes descriptifs d'objets) que le nombre d'attributs possédés par un même objet, dans la population étudiée, est sinon invariable, du moins de variance faible. *Cette circonstance confère aux méthodes basées sur la donnée d'une préordonnance associée à une similarité, un caractère intrinsèque, c'est-à-dire ne dépendant pas du choix de la mesure de similarité.*

2.6 — *Modèle probabiliste*

E peut être un échantillon d'une population E' plus vaste ; de toute façon E nous est donné comme un échantillon dans l'ensemble des parties de l'ensemble des attributs A, $P(A)$ (d'une manière équivalente dans $\{0, 1\}^T$). Dans ce cadre on peut regarder E comme un échantillon de n éléments aléatoires indépendants pris dans $P(A)$ muni d'une mesure de probabilité et chercher à déterminer l'espérance mathématique du nombre d'inversions que présente $\underline{\omega}$ par rapport à $\overline{\omega}$. Nous avons déjà vu que si cette mesure de probabilité est concentrée sur un seul niveau du simplexe $P(A)$ ou sur deux niveaux successifs, cette espérance mathématique ne peut être que nulle ; il faut aussi s'attendre à ce qu'elle soit petite si, pour cette mesure de probabilité, la variance de la variable aléatoire $|X|$ cardinal de X, $(X \subset A)$, est petite.

x, y, z et t etant un échantillon de 4 objets indépendants pris dans $P(A)$, désignons par

$$\Phi(\{x, y\}, \{z, t\})$$

la fonction indicatrice de l'événement

$$\left\{ \sum_i (x_i - y_i)^2 > \sum_i (z_i - t_i)^2 \ , \ \sum_i x_i y_i > \sum_i z_i t_i \right\} \tag{1}$$

qui est une inversion.

Le nombre d'inversions ν est égal à

$$\sum \Phi(\{x, y\}, \{z, t\}) + \sum \Phi(\{x, y\}, \{x, z\}) \tag{2}$$

la première somme étant étendue à G et la seconde à H.

$$E(\nu) = \sum E[\Phi(\{x, y\}, \{z, t\})] + \sum E[\Phi(\{x, y\}, \{x, z\})] \tag{3}$$

$$E[\Phi(\{x, y\}, \{z, t\})] = \Pr\left\{ \sum_i (x_i - y_i)^2 > \sum_i (z_i - t_i)^2 \ , \ \sum_i x_i y_i > \sum_i z_i t_i \right\} \tag{4}$$

$$E[\Phi(\{x, y\}, \{x, z\})] = \Pr\left\{ \sum_i (x_i - y_i)^2 > \sum_i (x_i - z_i)^2 \ , \ \sum_i x_i y_i > \sum_i x_i z_i \right\} \tag{5}$$

En notant π_1 l'expression (4) et π_2 l'expression (5), on a

$$E(\nu) = \frac{1}{4} n(n-1)(n-2)(n-3)\pi_1 + n(n-1)(n-2)\pi_2 \tag{6}$$

$$= n(n-1)(n-2)\left(\frac{n-3}{4}\pi_1 + \pi_2\right) .$$

Le nombre de couples de paires distinctes étant

$$w = \frac{1}{8}(n+1)n(n-1)(n-2) ,$$

$$E\left(\frac{\nu}{w}\right) = \frac{2(n-3)}{(n+1)}\pi_1 + \frac{8}{n+1}\pi_2 , \tag{7}$$

Si n est assez grand, $E(\nu/w)$ est approximativement égal à $2\pi_1$. Dans ces conditions, seul π_1 nous intéressera. Avec des notations plus synthétiques :

$$\pi_1 = \Pr\{u + v > u' + v' , s > s'\} \tag{8}$$

où $I(x, y) = (s, u, v)$ et $I(z, t) = (s', u', v')$.

Appelons B_k l'événement $\{s = k\}$, B_h celui $\{s' = h\}$, C_l, $\{u + v = l\}$ et C_m, $\{u' + v' = m\}$. On a

$$\pi_1 = \sum_{l>m} \sum_{k>h} \Pr\{C_l \cdot C_m \cdot B_k \cdot B_h\} , \tag{9}$$

les événements $C_l \cdot B_k$ et $C_m \cdot B_h$ sont indépendants ; le premier étant associé au couple (x, y) et le second à (z, t), d'où

$$\pi_1 = \sum_{l>m} \sum_{k>h} \Pr(B_k \,.\, C_l) \Pr(B_h \,.\, C_m) \;, \quad (10)$$

par conséquent dans la mesure où le modèle nous fournit les probabilités $p_{kl} = \Pr(B_k \,.\, C_l)$, π_1 peut être déterminé.

Exemple : Nous avons à *titre d'exercice* déterminé π_1 dans le cadre du modèle suivant : x étant un objet quelconque de E dont le représentant dans $\{0\,,\,1\}^T$ est $a(x) = (x_1, x_2, \ldots, x_l, \ldots, x_T)$; on suppose que pour tout i

$$\Pr\{x_i = 1\} = p \qquad 0 < p < 1$$

d'où $\Pr\{x_i = 0\} = 1 - p$;

de plus, l'indépendance en probabilité des différentes coordonnées est supposée. Soulignons que ce cas peut très difficilement correspondre à un exemple réel.

Déterminons les nombres $p_{kl} = \Pr(B_k \,.\, C_l)$.

$$\Pr(B_k \,.\, C_l) = \Pr(C_l/B_k) \Pr(B_k) \qquad (11)$$

$$\Pr(B_k) = \binom{T}{k} p^{2k} (1 - p^2)^{T-k}$$

et $$\Pr(C_l/B_k) = \binom{T-k}{l} \left(\frac{2p}{1+p}\right)^l \left(\frac{1-p}{1+p}\right)^{T-k-l}$$

D'où la table des valeurs de π_1.

	$T = 10$	$T = 20$	$T = 30$	$T = 40$
$p = 0{,}1$	0,029	0,058	0,081	0,099
$p = 0{,}2$	0,079	0,120	0,140	0,151
$p = 0{,}3$	0,099	0,128	0,141	0,150
$p = 0{,}4$	0,091	0,114	0,125	0,132
$p = 0{,}5$	0,072	0,092	0,102	0,108
$p = 0{,}6$	0,049	0,066	0,075	0,081
$p = 0{,}7$	0,026	0,046	0,048	0,054
$p = 0{,}8$	0,009	0,017	0,023	0,028
$p = 0{,}9$	0,001	0,002	0,004	0,006

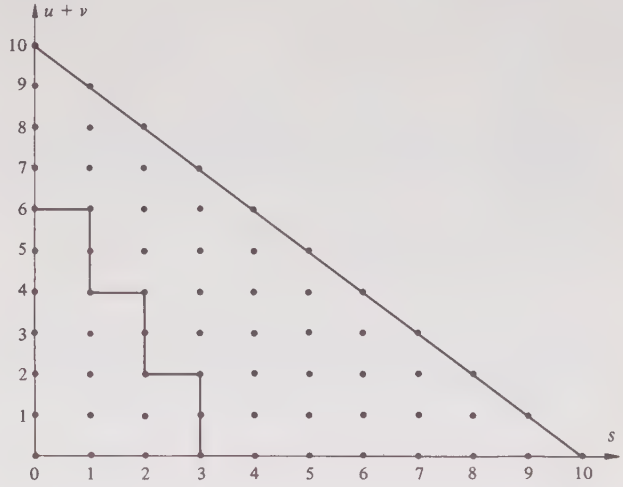
2.7 — *Calcul pratique du nombre d'inversions*

Concrètement, quand il s'agira de s'assurer de la stabilité de la préordonnance associée, on calculera l'écart, nombre d'inversions entre les deux préordres les plus éloignés que sont $\underline{\omega}$ et $\overline{\omega}$, définis sur E. Pour cela à chaque paire $\{x\,,\,y\}$ faisons correspondre le point de $N^2 \left(\sum_l x_l y_l\,,\, \sum_l (x_l - y_l)^2\right) = (s\,,\,u+v)$ et désignons par n_k^h $(h \leqslant T-k)$ le cardinal de l'ensemble

$$A_k^h = \{\{x\,,\,y\} \in F \mid s = k\,,\, u + v = h\}\,.$$

En affectant au point $(k\,,\,h)$ la fréquence n_k^h, nous définissons la distribution empirique sur F du couple $(s\,,\,u+v)$.

Représentation graphique $(T = 10)$



Ordonnons par valeurs croissantes de k les classes

$$B_k = \{\{x\,,\,y\} \in F \mid s = k\} = \sum_{h=0}^{T-k} A_k^h$$

(somme ensembliste).

Le nombre d'inversions que présentent les éléments de A_k^h avec les éléments des différentes classes B_j précédant B_k, $(j < k)$, est

$$n_k^h \sum_{l=0}^{k-1} \sum_{m=0}^{k-1} n_l^m \; ; \qquad (1)$$

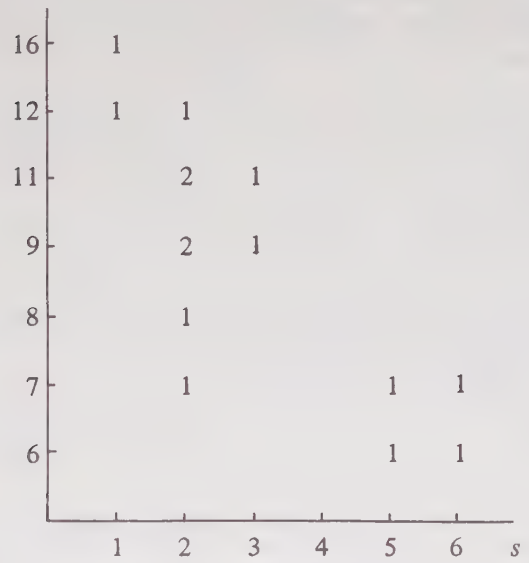
la sommation est étendue au rectangle de points :

$$0 \leqslant l \leqslant (k-1) \quad , \quad 0 \leqslant m \leqslant (h-1)$$

d'où le nombre d'inversions

$$v = \sum_{k=0}^{T} \sum_{h=0}^{T-k} n_k^h \sum_{l=0}^{k-1} \sum_{m=0}^{h-1} n_l^m \; . \qquad (2)$$

Exemple : Reprenons l'exemple du paragraphe 1 ci-dessus.



$$v = 2 + 4 + 1 = 7$$

Il résulte de la représentation graphique ci-dessus que la condition nécessaire et suffisante pour qu'un ensemble d'objets soit tel que $v = 0$, est qu'il existe une fonction de N dans N, non croissante, pour laquelle, les points $(k\,,\,h)$ du triangle de

$$N^2 \; (k + h \leqslant T) \; ,$$

de fréquence n_k^h non nulle, appartiennent à son graphe.

IV — *PROPOSITION DE QUELQUES PREORDONNANCES PARTICULIERES SUR E*

Nous allons néanmoins, dans le cas où on ne peut considérer la variance V comme tout à fait négligeable, commencer par proposer deux préordonnances dont chacune est adaptée à une situation particulière. Pour des raisons qui apparaîtront bientôt, nous appellerons la première la préordonnance la "plus fine" et la seconde la "moins fine".

1 — La préordonnance la "plus fine"

Dans le cadre des conditions générales [§ II 1.1], on suppose de plus que la prise en compte des associations négatives peut être négligée. Cette situation peut être illustrée quand, par exemple, les conditions suivantes semblent astisfaites :

Les différents caractères sont bivalents, les deux modalités d'un même caractère sont telles que

a) également significatives de la ressemblance ou bien,

b) l'une des deux modalités est significative de la ressemblance alors que l'autre ne l'est pas.

On définira l'ensemble A des attributs en retenant pour chacun des caractères la ou les deux modalités significatives.

1.1 — *Définition de* ω^1. Etant donné deux couples quelconques $(x\,,\,y)$ et $(z\,,\,t)$ auxquels sont associés leurs indicateurs respectifs

$$I(x\,,\,y) = (s\,,\,u\,,\,v) \text{ et } I(z\,,\,t) = (s'\,,\,u'\,,\,v') \; .$$

nous poserons

$$\{x\,,\,y\} \leqslant \{z\,,\,t\} \begin{cases} s > s' \\ \text{ou} \\ s = s' \text{ et } u + v < u' + v' \\ \text{ou} \\ s = s',\, u + v = u' + v' \text{ et } uv \leqslant u'v' \end{cases}$$

En d'autres termes

$$\{x\,,\,y\} \leqslant \{z\,,\,t\} \Longleftrightarrow (T - s\,,\,u + v\,,\,uv) \leqslant (T - s'\,,\,u' + v'\,,\,u'v')\,,$$

où la relation du second membre est celle de l'ordre sur N^3, produit lexicographique des ordres définis respectivement sur l'ensemble N des entiers naturels.

La préordonnance définie a un caractère central, par rapport à celles associées à des mesures de similarité, que nous expliciterons par la proposition suivante.

Désignons toujours par $\omega(S)$ la préordonnance associée à une similarité $S(x\,,\,y)$ et par $\equiv$ (resp. $\equiv_S$) l'équivalence sur F associée au préordre ω^1 (resp. $\omega(S)$) ; (rappelons que cette équivalence est définie de la façon suivante : $p \equiv$ (resp. $\equiv_S$) $q \Longleftrightarrow p \leqslant q$ et $q \leqslant q$ pour ω^1 (resp. $\omega(S)$) où $p\,,\,q \in F$).

1.2 — *Proposition.* L'équivalence $\equiv$ est plus fine que $\equiv_S$ associée à une mesure de similarité quelconque.

En effet, si $\{x\,,\,y\} \equiv \{z\,,\,t\}$, c'est que $s = s'$, $u + v = u' + v'$ et $uv = u'v'$; donc $u = u'\,,\,v = v'$ et par suite $S(x\,,\,y) = S(z\,,\,t)$ donc $\{x\,,\,y\} \equiv_S \{z\,,\,t\}$.

Lorsque les hypothèses que nous avons évoquées avant la définition de ω^1 sont satisfaites, l'habitude consistait à considérer la préordonnance $\omega(S)$, associée à une similarité fonction de s seulement. La préordonnance proposée tient compte de plus près, d'une manière naturelle, de l'information contenue dans l'indicateur $(s\,,\,u\,,\,v)$ d'un couple $(x\,,\,y)\,,\,(x\,,\,y)$ variant dans $E \times E$. La préordonnance ω^1 est évidemment compatible avec $\omega(S)$ où S est fonction de s; compte tenu de la proposition ci-dessus, ω^1 est en général sensiblement plus fine que $\omega(S)$.

Exemple : Définissons ω^1 dans le cas de l'exemple (§ II.2) ; on a

$$\{a\,,f\} < \{a\,,d\} < \{d\,,f\} < \{b\,,c\} < \{e\,,f\} < \{a\,,c\} < \{b\,,c\} < \{d\,,e\} < \{c\,,e\} < \{b\,,d\} < \{a\,,b\} < \{c\,,d\} < \{b\,,f\} < \{a\,,e\} < \{c\,,f\}\ .$$

La préordonnance "la plus fine" est ici une ordonnance.

2 — La préordonnance la "moins fine"

Toujours dans le cadre des hypothèses générales [§ II. 1. 1], on ne veut rien postuler sur la contribution à la mesure de la ressemblance d'une association négative par rapport à celle d'une association positive. Pour cette situation on préfèrera un préordre, peut-être grossier, mais sûr.

2.1 — *Définition de* ω^2. $(s\,,u + v)$ et $(s'\,,u' + v')$ étant respectivement associés aux paires $p = \{x\,,\,y\}$ et $q = \{z\,,\,t\}$, on posera $p \leqslant q$ si et seulement si $s \geqslant s'$ ou $u + v \leqslant u' + v'$.

De la sorte on n'a $p > q$ que si $s < s'$ et

$$u + v > u' + v'\ .$$

Le préordre ω^2 a encore un caractère central, par rapport aux préordonnances associées aux similarités de la forme $\mathfrak{S}(s\,,\,u + v)$, que nous exprimerons par la

2.2 — *Proposition*

La préordonnance ω^2 est moins fine que toute préordonnance associée à une mesure de similarité de la forme $\mathfrak{S}(s\,,\,u + v)$.

Il est, en effet, manifeste que si on a strictement

$$\{x\,,\,y\} < \{z\,,\,t\} \quad \text{pour} \quad \omega^2$$

alors $\mathfrak{S}(s\,,\,u + v) < \mathfrak{S}(s'\,,\,u' + v')$ où $(s\,,\,u + v)$ et $(s'\,,\,u' + v')$ sont respectivement associés à $\{x\,,\,y\}$ et à $\{z\,,\,t\}$.

2.3 — *Remarque*

Le choix de ω^1 ou de ω^2 illustre le caractère arbitraire souligné au paragraphe III.1. d'une préordonnance associée à une similarité qui soit précisément une certaine *fraction rationnelle* en s, $u + v$ et uv.

Exemple : Définissons ω^2 dans le cas de l'exemple (§ II.2) ; on a

$$\{a\,,f\}=\{a\,,d\}=\{d\,,f\}=\{b\,,c\}=\{b\,,e\}<\{a\,,c\}$$
$$=\{e\,,f\}=\{a\,,b\}=\{b\,,d\}=\{c\,,d\}=\{c\,,e\}$$
$$=\{d\,,e\}=\{b\,,f\}=\{a\,,e\}<\{c\,,f\}\,.$$

La préordonnance "la moins fine" comporte ici 3 classes.

3 — Préordonnances associées à des indices centrés et réduits

Il reste que, lorsque la variance V, du nombre d'attributs possédés par un même objet, est assez grande, (l'expérience seule pouvant préciser ce dernier qualificatif) , il est important de se méfier des effets statistiques qui feront que deux objets de grosse taille (i.e. ayant chacun un grand nombre d'attributs) auront tendance à être considérés, à priori, davantage ressemblants que deux objets de petite taille, du point de vue de l'un quelconque des indices de similarité $\mathcal{S}\,(s\,,u\,,v)$. Il y a lieu par conséquent de corriger le biais qui apparaît pour des indices définis à partir de s ou de $(s+t)$.

Si X et Y sont les deux parties de l'ensemble A des attributs, que définissent deux objets x et y, pour lesquelles on a $|X|=l_x$ et $|Y|=l_y$, P. Achard nous suggère de rapporter la valeur de $s=|X\cap Y|$ à celle du "hasard parfait" où X (resp. Y) serait pris dans l'ensemble des parties de A à l_x (resp. l_y) éléments ; chacun de ces deux ensembles étant muni d'une probabilité uniformément répartie.

Dans le cadre d'une telle hypothèse $|X\cap Y|$ est une variable binomiale de paramètres $(T\,,p)$ où $p=l_x\,.\,l_y/T^2$. Par conséquent, la statistique qui exprime une mesure sans biais de la similarité des deux objets x et y, lorsque seule la présence commune d'un attribut est significative de la ressemblance, est

$$(s-Tp)/\sqrt{Tp\,(1-p)}$$

soit, en introduisant les paramètres s, u et v

$$\frac{(s-(s+u)\,(s+v)/T)}{\sqrt{((s+u)\,(s+v)/T)\,(1-(s+u)\,(s+v)/T^2)}}$$

D'autre part, pour le cas où la présence commune et l'absence commune d'un même attribut chez deux objets peuvent être considérées comme également significatives de la ressemblance , on corrigera, lorsque la variance V est grande, l'indice $(s+t)$, le centrant et le réduisant par rapport à l'hypothèse exprimée ci-dessus. On obtient

$$((s+t)-Tq)/\sqrt{Tq\,(1-q)}$$

où

$$q=(l_x l_y+(T-l_x)\,(T-l_y))/T^2=$$
$$=((s+u)\,(s+v)+(t+v)\,(t+u))/T^2.$$

CHAPITRE II

CRITÈRES

I — *INTRODUCTION*

On distingue, en général, pour une classification donnée deux concepts de classe. Une classe est dite monothétique si elle est caractérisée par une propriété, ce qui signifie qu'un objet appartient à la classe si et seulement s'il possède la propriété. Une telle définition de classe est trop rigide lorsqu'on se réfère aux classifications naturelles dans les Sciences humaines ; comme le notait déjà Vicq d'Azyr (1792), une classe peut être très "naturelle" sans qu'il soit possible de définir un caractère commun à tous les éléments la composant. On oppose à cette dernière notion de classe celle de classe polythétique précisée de la façon suivante par Beckner (1959) : une classe K se réfère à un ensemble J d'attributs : $a_1, a_2, \ldots, a_j$ tels que

a) chaque élément de la classe possède une proportion importante (mais non fixée) d'attributs de J ;

b) chaque attribut de J est possédé par une proportion importante d'éléments de K ;

c) il n'y a pas nécessairement un attribut de J qui soit possédé par tous les éléments de K.

En considérant pour chaque couple d'objet son indicateur (s, u, v), précédemment défini, ou mieux la préordonnance sur E associée au choix d'une similarité, et en se référant à la notion de classe polythétique il convient que la ressemblance de deux objets soit assez grande lorsqu'ils appartiennent à la même classe et suffisamment petite dans le cas contraire.

Dès lors, on cherche sur l'ensemble des objets à découvrir soit une classification (i.e. une partition), soit une hiérarchie de classifications (i.e. une chaîne dans le treillis des partitions) qui respecte d'une manière satisfaisante les ressemblances entre objets. Bien que la nature des techniques utilisées soit la même pour les deux problèmes (recherche d'une classification ou d'une hiérarchie de classifications) le second ne se ramène pas mathématiquement au premier. C'est pour cette raison que nous distinguerons dans la suite les deux buts.

La structure σ à laquelle appartient la donnée de base est moins "riche" que celle, τ, de l'objet recherché (partition ou chaîne de partitions). La démarche consiste à injecter la classe Θ, des structures, τ, sur E dans celles, Σ des structures, σ, sur E. Il s'agit alors de déterminer celui ou ceux des éléments de la classe Θ des structures τ qui sont au mieux compatibles avec la donnée de base. Pour cela on définit un *critère* ; un critère est, très généralement, une relation de préordre (qui peut être partielle) sur l'ensemble Θ. Pour cette relation τ est préférable à τ' ($\tau' < \tau$) si, en un sens qu'il y a lieu de préciser, τ est plus proche que τ' de la structure définie par la donnée de base.

Le critère est le plus souvent défini à partir de la donnée d'une métrique ou pseudométrique sur Σ, mais nous verrons des cas où il n'en est pas ainsi.

La donnée de base , celle à partir de laquelle on travaillera peut être :

1/ Le tableau $A \times E$ (où A est l'ensemble des attributs, E celui des objets ; $|A| = T$ et $|E| = n$).

Dans ce cas Σ est l'ensemble des matrices d'incidence à T lignes et n colonnes (une matrice d'incidence est une matrice de zéros et de uns, indiquant une famille finie de sous-ensembles d'un ensemble fini ; ici chaque objet est considéré comme une partie de l'ensemble des attributs).

2/ Une relation binaire symétrique sur E, Σ est alors l'ensemble des relations binaires symétriques sur E.

3/ Une famille finie de p partitions sur E, Σ est dans ces conditions l'ensemble des familles finies de p partitions sur E.

4/ Une distance ou pseudo-distance sur E (indice de similarité), Σ est ici l'ensemble des pseudo-métriques sur E.

5/ Enfin une préordonnance sur E, Σ est alors l'ensemble des préordres totaux sur l'ensemble F des paires d'objets distincts.

Les différents paragraphes de ce chapitre se rapporteront respectivement à ces différents cas. Quel que soit cet espace, il sera possible de définir différents critères qui semblent intuitivement "naturels". De ce fait on sera conduit à entreprendre une étude comparée de quelques-uns des critères qui seront proposés. Le problème de la comparaison des critères est, à notre avis, du plus grand intérêt pour la classification automatique. Dans son état actuel, cette comparaison reste insuffisante. Pour cette raison, le traitement proposé du sujet est loin d'être homogène et systématique. Il reste néanmoins que ce chapitre se situe au cœur de notre travail de synthèse.

II — *CRITERES DEFINIS A PARTIR DE LA MATRICE D'INCIDENCE DES DONNEES*

1 — Généralités

Rappelons que le tableau des données se présente sous la forme d'une matrice d'incidence (ϵ_{ij}), $i = 1, 2, \ldots, T$ et $j = 1, 2, \ldots, n$; où $\epsilon_{ij} = 1$ si l'attribut a_i est présent chez l'objet codé j et 0 sinon. Chaque objet étant représenté par une colonne de la matrice, une partition de l'ensemble des objets sera représentée par une partition des colonnes. On convient dans le tableau de regrouper les colonnes relatives à une même classe.

Soit $(E_1, E_2, \ldots, E_k)$ une partition de E en k classes ; si cette partition définit une classification naturelle, étant donné une classe E_i, il lui correspond, en citant Beckner, un sous-ensemble A_i des attributs, auquel elle se réfère, tel que chaque caractère de A_i est fréquent dans E_i. Une telle classe sera d'autant plus cohérente que la fréquence des uns, dans la restriction de là matrice au tableau $A_i \times E_i$ (respectivement au rectangle $A_i \times (\sum_{j \neq i} E_i)$) est grande (respectivement petite).

Les sous-ensembles A_i ne forment pas nécessairement une partition mais généralement un recouvrement, car un même attribut peut intervenir dans la définition de plus d'une classe. Ce recouvrement est assez fin, nous voulons dire que les parties A_i sont faiblement empiétantes ; (pour $i \neq j$, le rapport $|A_i \cap A_j| / |A_i \cup A_j|$ est petit s'il n'est pas nul). En effet si pour (i, j) donné, $i \neq j$, le rapport

$$|A_i \cap A_j| / |A_i \cup A_j|$$

est assez grand, il n'y a aucune raison à partir de $E_i \cup E_j$ de constituer deux classes.

2 — Définition d'un critère

L'objet de notre recherche est donc un couple dont le premier argument est une partition sur E et le second une famille de parties de A, recouvrant A, telle que les éléments de la famille et les classes de la partition se correspondent bijectivement. Avec des notations que l'on comprend, désignons par $(E_i, A_i / i \in I)$ un tel couple.

Une technique directe de partition opérerait à partir de la matrice d'incidence par permutation des lignes ou des colonnes. Mais il semble difficile à ce niveau de définir un critère proprement conçu ; un tel critère dépendrait, comme nous venons de le voir, d'un couple d'arguments du type $(E_i, A_i / i \in I)$ ci-dessus.

Ce critère devrait avantager ceux des $(E_i, A_i/i \in I)$ pour lesquels la fréquence des uns est importante dans la restriction du tableau à $\bigcup_I A_i \times E_i$ et faible dans sa restriction au complémentaire de $\bigcup_I A_i \times E_i$; on voudrait d'autre part que la famille $(A_i)_{i \in I}$ soit faiblement empiétante.

Nous exposerons deux types de critères, chaque type étant défini à partir d'une forme de restriction à la généralité du problème.

Nous désignerons dans la suite de ce paragraphe par L (resp. par M) la restriction de la matrice d'incidence à $\bigcup_{i \in I} A_i \times E_i$ (resp. au complémentaire de $\bigcup_{i \in I} A_i \times E_i$).

2.1 — Représentation du couple $(E_i, A_i/i \in I)$ au moyen d'une matrice de zéros et de uns.

• La matrice Π de zéros et de uns représentant $(E_i, A_i/i \in I)$ est définie comme suit :

$$\begin{aligned} \varpi_{ij} &= 1 \quad \text{si} \quad (i, j) \in L \\ &= 0 \quad \text{si} \quad (i, j) \in M = L^c \ . \end{aligned}$$

Usant le langage du sens commun, une telle matrice est relative au cas idéal où la classification $(E_i)\, i \in I$ s'impose d'une manière évidente ; une classe E_i étant caractérisée par un sous-ensemble d'attributs A_i dans le sens suivant : un objet appartient à la classe E_i si et seulement s'il possède exactement le sous-ensemble d'attributs A_i.

2.2 — On peut imposer à la famille $(A_i)_{i \in I}$ d'être une partition sur A et cela au risque d'affecter à une seule classe un attribut relatif à plus d'une. Malgré cette restriction, on espère, une fois la partition obtenue, déterminer les caractères les plus spécifiques d'une classe donnée.

Soit $(E_i, A_i | i \in I)$ un couple formé par une partition sur E et une autre sur A se correspondant bijectivement. Entre deux couples tels que ce dernier, on préférera celui pour lequel la valeur de

$$\Sigma\, (\epsilon_{ij} - \varpi_{ij})^2 \qquad (1)$$

est la plus petite où $\Pi = (\varpi_{ij})$ est la matrice de zéros et de uns relative à $(E_i, A_i | i \in I)$.

La quantité ci-dessus représente la restriction de la distance euclidienne définie dans $R^{T \times n}$ à $\{0, 1\}^{T \times n}$, entre le vecteur représentant la matrice d'incidence des données et celui relatif à Π.

Associant à une matrice d'incidence la relation R définie sur $A \cup E$ de la façon suivante :

$$x\, R\, y \iff \begin{cases} (1)\ x \in A \quad \text{et} \quad y \in E \\ \text{et} \\ (2)\ \text{l'attribut } x \text{ est présent dans } y\ , \end{cases}$$

la distance ci-dessus représente le cardinal de la différence symétrique entre les deux graphes des relations définies respectivement par la matrice d'incidence des données et Π dans $A \cup E \times A \cup E$.

Notons que cette distance peut se mettre sous la forme

$$\sum_M \epsilon_{ij} - \sum_L \omega_{ij} + \sum_{i \in I} |A_i| \times |E_i| \ .$$

Dans ce cadre, J.P. Benzecri proposerait de maximiser $\sum_I \frac{\nu_i^2}{|A_i \times E_i|}$ où ν_i représente le nombre de uns dans le sens tableau $A_i \times E_i$.

2.3 — On peut imposer le nombre de classes, k, de la partition définie sur E. Dans cette optique on peut reprendre en l'aménageant le critère que nous avons proposé ci-dessus.

Dans la suite nous préciserons les exigences que nous avons exprimées intuitivement dans l'introduction pour exposer les différents critères permettant de déterminer une classification (resp. une hiérarchie de classification) sur E qui respecte au mieux les ressemblances entre objets. Une fois la classification obtenue, on se préoccupera de définir relativement à une même classe ses attributs dominants.

Exemple : Le tableau ci-dessous représente le tableau d'incidence décrit au chapitre 1 (§ II.2) où les objets se présentent dans un ordre différent et où, pour respecter l'usage, les rôles des lignes et des colonnes ont été inversés.

	a	d	f	e	b	c
α_1	0	0	1	1	0	0
α_2	1	1	0	0	0	0
α_3	0	0	0	1	0	1
α_4	1	1	1	0	0	0
α_5	1	0	1	0	0	0
α_6	0	1	1	1	0	0
α_7	0	0	0	0	1	1
α_8	1	0	1	0	0	0
α_9	0	0	1	0	0	0
α_{10}	1	0	0	1	1	1
α_{11}	1	1	1	0	0	1
α_{12}	0	0	0	0	0	1
α_{13}	1	1	0	0	0	1
α_{14}	0	1	1	1	1	0
α_{15}	0	0	0	0	1	1
α_{16}	1	0	1	0	0	0
α_{17}	1	1	1	0	1	0
α_{18}	0	0	0	0	1	1

Nous allons comparer du point de vue du critère (1) ci-dessus les deux partitions en deux classes

$$(\{a, d, f\} \quad ,\{b, c, e\}) \qquad \text{(I)}$$

et

$$(\{a, d, e, f\}, \{b, c\}) \qquad \text{(II)}$$

Examinons (I) ; dans le cas 2.2 précédent il y a lieu de définir sur l'ensemble A des attributs une partition en deux classes (A_1, A_2) correspondant à (I). Les deux classes de (I) ayant le même nombre d'éléments, une procédure optimale consiste à affecter à A_1 les attributs qui sont strictement plus fréquents dans la classe $\{a, d, f\}$. Cet algorithme se généralise à un nombre quelconque de classes : on affectera l'attribut α_i à la classe de numéro k pour laquelle

$$\sum_j (\epsilon_{ij} - \varpi_{ij})^2 ,$$

où $\varpi_{ij} = 1$ si $j = k$ et 0 sinon, est minimum.

On a

$$(A_1, A_2) = (\{\alpha_2, \alpha_4, \alpha_5, \alpha_6, \alpha_8, \alpha_9, \alpha_{11}, \alpha_{13}, \alpha_{16}, \alpha_{17}\}, \{\alpha_1, \alpha_3, \alpha_7, \alpha_{10}, \alpha_{12}, \alpha_{14}, \alpha_{15}, \alpha_{18}\}).$$

En sommant ligne après ligne, on obtient pour la valeur du critère (1)

$$3+1+1+0+1+2+1+1+2+1+2+2+2+2+1+1+1+1=25$$

Dans le cas 2.3 ci-dessus, il y a lieu de définir un couple de parties (A'_1, A'_2), A'_1 correspondant à la classe $\{a, d, f\}$ et A'_2 à $\{b, c, e\}$. Ici une procédure naturelle consiste à attacher un attribut donné aux classes où il est le plus fréquent, au sens large ; pour cela on calculera pour chacune des classes le rapport n_k^i/n_k, où n_k est le cardinal de la classe et n_k^i le nombre de fois où l'attribut α_i est présent dans la classe de numéro k ; et on affectera α_i à celles des classes pour lesquelles ce rapport est le plus grand. Ainsi

$$(A'_1, A'_2) = (\{\alpha_1, \alpha_2, \alpha_4, \alpha_5, \alpha_6, \alpha_8, \alpha_9, \alpha_{11}, \alpha_{13}, \alpha_{14}, \alpha_{16}, \alpha_{17}\} , \{\alpha_1, \alpha_3, \alpha_7, \alpha_{10}, \alpha_{12}, \alpha_{14}, \alpha_{15}, \alpha_{18}\}).$$

La valeur du critère (1) est dans ces conditions

$$4+1+1+0+1+2+1+1+2+1+2+2+2+2+1+1+1+1=26$$

Examinons à présent II ; dans le cas 2.2

$$(A_1, A_2) = (\{\alpha_1, \alpha_2, \alpha_4, \alpha_5, \alpha_6, \alpha_8, \alpha_9, \alpha_{11}, \alpha_{14}, \alpha_{16}, \alpha_{17}\} , \{\alpha_3, \alpha_7, \alpha_{10}, \alpha_{12}, \alpha_{13}, \alpha_{15}, \alpha_{18}\}) .$$

La valeur du critère (1) ci-dessus est

$$2+2+2+1+2+1+0+2+3+2+2+1+3 \\ +2+0+2+2+0=29$$

$$(A_1', A_2') =$$

$$(\{\alpha_1, \alpha_2, \alpha_4, \alpha_5, \alpha_6, \alpha_8, \alpha_9, \alpha_{11}, \alpha_{13}, \alpha_{14}, \alpha_{16}, \alpha_{17}\}, \\ \{\alpha_3, \alpha_7, \alpha_{10}, \alpha_{12}, \alpha_{13}, \alpha_{15}, \alpha_{18}\}).$$

La valeur du critère (1) est encore 29.

La partition (I) est donc meilleure que (2) au sens du critère (1).

III — *CRITERES DEFINIS A PARTIR D'UNE RELATION BINAIRE SYMETRIQUE SUR E*

Une telle relation est supposée en général non réflexive.

1 — Origine de la relation

Elle peut être définie à partir du choix d'un indice de similarité de la façon suivante

$$x\,R\,y \iff S(x, y) \geqslant s$$

où s est un seuil donné. En termes de préordonnance on a $x\,R\,y$ et si seulement si la paire $\{x, y\}$ se trouve à gauche d'une borne donnée. En prenant $S(x, y)$ sous la forme $\mathcal{S}(s, u+v)$, une telle donnée est relativement stable si la variance du nombre d'attributs possédés par un même objet est assez petite.

2 — Représentation d'une partition

La donnée d'une partition sera ici représentée par celle d'une relation d'équivalence sur E, c'est-à-dire une relation binaire réflexive *symétrique* et transitive.

3 — Définition d'un critère

Dans un article paru en 1964, C.T. Zahn [39] étudiait le problème de trouver une relation d'équivalence Π qui approxime au mieux une relation symétrique donnée R dans le sens suivant : "Minimiser le cardinal de la différence symétrique des graphes de R et de Π dans $E \times E$". L'auteur résout le problème pour une classe de relations R qui ne peut nous intéresser ici.

Mais nous avons là un critère qui préordonne totalement l'ensemble des partitions selon la valeur du cardinal de la différence symétrique entre la relation d'équivalence relative à la partition et R.

$(n_1, n_2, \ldots, n_k)$ désignant le type d'une partition P, M. Barbut exprime $|R \,\Delta\, \Pi|$ de la façon suivante

$$|R \,\Delta\, \Pi| = |R| + |\Pi| - 2\,|R \cap \Pi| \qquad (1)$$

$$= 2\,|R - R \cap \Pi| - |R| + |\Pi| \qquad (2)$$

Relativement à deux classes E_i et E_j de cardinaux respectifs n_i et n_j, on introduit le cardinal r_{ij} de l'ensemble des paires de la forme $\{x, y\}$ tels que x appartient à E_i, y appartient à E_j et $x\,R\,y$. Dans ces conditions (2) s'exprime sous la forme

$$4 \sum_{i<j} r_{ij} + \sum_i n_i(n_i - 1) - |R| \,. \qquad (3)$$

Ajoutant et retranchant $2 \sum_{i<j} n_i n_j$ on obtient

$$4 \sum_{i<j} \left(r_{ij} - \frac{n_i n_j}{2}\right) + n^2 - n - |R| \quad . \qquad (4)$$

Posant $d_{ij} = r_{ij} - \dfrac{n_i n_j}{2}$, le critère préordonne l'ensemble des partitions selon la valeur de

$$\sum_{i<j} d_{ij} \,. \qquad (5)$$

Les relations R et Π étant symétriques, nous nous exprimerons ci-dessous dans l'ensemble F des paires

d'objets distincts de E, quotient de $E \times E$ par la relation $(x, y) \equiv (y, x)$ pour tout couple (x, y) de $E \times E$. De la sorte, la donnée d'une relation symétrique sur E est équivalente à la donnée d'une partie de F.

Exemple : Soit $E = \{a, b, c, d, e, f\}$ et

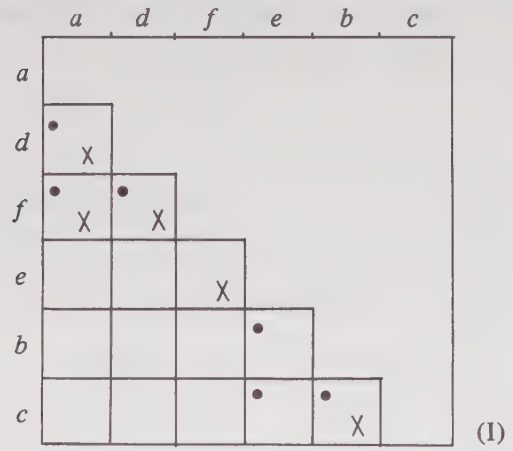
$$R = \{\{a, d\}, \{a, f\}, \{b, c\}, \{d, f\}, \{e, f\}\}.$$

Considérons les deux partitions

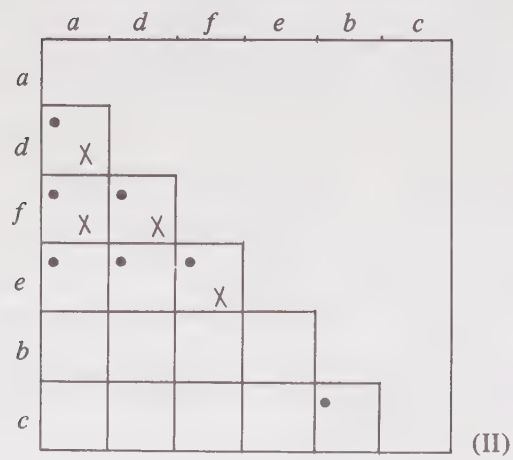
$$\{\{a, d, f\}, \{b, c, e\}\} \quad \text{(I)}$$

$$\{\{a, d, e, f\}, \{b, c\}\} \quad \text{(II)}$$

dont nous représentons les graphes dans les tableaux suivants par les points.



(I)



(II)

Le graphe de R est figuré par les croix. Pour (I) $|R \Delta \Pi| = 3$ et pour (II) $|R \Delta \Pi| = 2$. Donc la partition (II) est préférable à (I) pour ce critère.

4 — Aspect limite du critère défini

J.W. Moon [20], se référant à l'article de Zahn, établit un résultat limite dont la généralité le rendra intéressant pour la classification automatique. Nous allons reprendre cette étude et montrer son impact sur nos problèmes.

• R étant fixée, nous désignerons par $m(R)$ la valeur du minimum du cardinal de la différence symétrique entre R et une relation d'équivalence Π.

$$m(R) = \min_{\Pi \in \mathcal{R}(\Pi)} |R \Delta \Pi|$$

où $\mathcal{R}(\Pi)$ est l'ensemble des relations d'équivalence sur E ; R et Π sont considérées ici comme deux parties de F (cf. ci-dessus).

4.1 — *Lemme* : On a $m(R) \leqslant \frac{1}{2}\binom{n}{2}$

Soit $r = |R|$; si r est inférieur ou égal à $\frac{1}{2}\binom{n}{2}$, en prenant pour Π la relation d'équivalence relative à la partition la plus fine (chaque classe contient un objet), on a $|R \Delta \Pi| \leqslant \frac{1}{2}\binom{n}{2}$ et par suite $m(R) \leqslant \frac{1}{2}\binom{n}{2}$. Si r est strictement supérieur à $\frac{1}{2}\binom{n}{2}$, en prenant pour Π la relation d'équivalence associée à la partition la moins fine (tous les objets forment une seule classe) on a $|R \Delta \Pi| \leqslant \frac{1}{2}\binom{n}{2}$ et par suite $m(R) \leqslant \frac{1}{2}\binom{n}{2}$ d'où le résultat annoncé.

4.2 — *Théorème* : Pour tout ϵ réel positif,

$$\frac{1-\epsilon}{2}\binom{n}{2} \leqslant m(R) \leqslant \frac{1}{2}\binom{n}{2}$$

pour presque toutes les relations R. (i.e., pour toutes les relations sauf pour une fraction tendant vers zéro lorsque n tend vers l'infini).

Preuve : Soit $f = \binom{n}{2} = |F|$ et soit G_n le cardinal de l'ensemble des relations d'équivalence pouvant être définies sur E. On sait bien que $G_n \leqslant n^n$ pour tout entier positif n.

Commençons par majorer le cardinal de l'ensemble des relations R pour lesquelles $m(R) = t$. Pour cela à chaque relation d'équivalence Π faisons correspondre l'ensemble des relations R qui sont à une distance t de Π ; chacune de ces relations est définie par le choix d'un sous-ensemble $\overline{t}$ de F de cardinal t, de la façon suivante :

Si p est un élément quelconque de $\overline{t}$

$p \in R$ si $p \notin \Pi$ et $p \notin R$ si $p \in \Pi$.

Si q est un élément quelconque de $F - \overline{t}$

$q \in R$ si $q \in \Pi$ et $q \notin R$ si $q \notin \Pi$.

Il en résulte que nous avons certainement

$$|\{R \mid m(R) = t\}| \leqslant G_n \binom{f}{t} \leqslant n^n \binom{f}{t}. \quad (1)$$

D'où un majorant du nombre de relations R pour lesquelles $m(R) < (1 - \epsilon)\, f/2$ est donné par

$$\sum_{t < (1-\epsilon)f/2} n^n \binom{f}{t} = n^n \sum_{t < (1-\epsilon)f/2} \binom{f}{t}$$

$$< n^n f \binom{f}{[(1-\epsilon)\, f/2]}. \quad (2)$$

Le cardinal de l'ensemble des relations R étant 2^f, la proportion de celles pour lesquelles

$$m(R) < (1 - \epsilon)\, f/2$$

est strictement inférieur à

$$n^n f \binom{f}{[(1-\epsilon)\, f/2]}\, 2^f \quad (3)$$

ce rapport étant encore strictement inférieur à

$$n^n f \binom{f}{[(1-\epsilon)\, f/2]} \binom{f}{[f/2]} \quad (4)$$

qui se met sous la forme

$$n^n \frac{[f/2]\,([f/2]-1) \ldots\ldots\ldots ([(1-\epsilon)\, f/2]+1)}{(f-[(1-\epsilon)\, f/2])\,(f-[(1-\epsilon)\, f/2]-1) \ldots (f-[f/2]+1)}$$

$$< n^n f (1+\epsilon)^{-\epsilon f/2} .$$

— Cette dernière inégalité, qui résulte d'un calcul élémentaire, permet d'établir que (3) tend vers zéro lorsque n tend vers l'infini.

4.3 — *Conclusion.* Pour n grand la distribution de $m(R)$ est donc très concentrée immédiatement à gauche de $\frac{1}{2}\binom{n}{2}$. Relativement à une relation symétrique R, (dont nous avons explicité l'origine [cf. § III.1]), une relation d'équivalence Π, (obtenue en l'occurrence au moyen d'un algorithme de classification) dont la distance à R est sensiblement inférieure à $\frac{1}{2}\binom{n}{2}$, peut être considérée comme une très bonne approximation. Dans ces conditions, si $|R| < \frac{1}{2}\binom{n}{2}$ $\left(\text{resp. } |R| > \frac{1}{2}\binom{n}{2}\right)$, la relation d'équivalence Π relative à la partition la plus fine (resp. la moins fine) est, pourvu que n soit assez grand, une très bonne approximation de R au sens du critère $|R \,\Delta\, \Pi|$; nous commenterons ce point dans la conclusion.

IV — *CRITERES DEFINIS A PARTIR DE LA DONNEE D'UNE FAMILLE FINIE DE PARTITIONS*

Nous reprenons ici les travaux de S. Régnier qu'on peut voir développés dans [26].

1 — Origine de la donnée

Chacun des objets de E étant décrit par un certain nombre fini de caractères qui peuvent correspondre

à divers "points de vue", la description peut être définie, comme nous l'avons vu, par une application :

$$E \xrightarrow{a} A_1 \times A_2 \times \ldots \times A_h \times \ldots \times A_p = \mathcal{A}$$

où A_h est l'ensemble des valeurs du caractère a_h et

$$a = (a_1, a_2, \ldots, a_h, \ldots, a_p) .$$

Les ensembles A_h sont supposés finis et n'avoir aucune métrique naturelle ; de la sorte chaque caractère a_h définit par lui-même une partition de l'ensemble E des objets où chaque classe est formée des objets identiques du point de vue a_h :

x et y dans la même classe $\Longleftrightarrow a_h(x) = a_h(y)$.

— Dans ces conditions les p caractères définissent p partitions, en général différentes, de l'ensemble E, [cf. Ch. 1 — § I.1].

2 — Représentation d'une partition

La donnée d'une partition P sera représentée par celle d'une famille finie de p partitions identiques à P.

3 — Définition d'un critère

3.1 — *Partition centrale.* Adoptons sur l'ensemble $\mathcal{P}$ des partitions sur E la distance d définie comme suit : si P et P' sont deux éléments de $\mathcal{P}$, $d(P, P')$ est le cardinal de la différence symétrique des graphes dans $E \times E$ associés respectivement aux relations d'équivalence que définissent P et P'.

Une telle distance induit sur l'ensemble S des familles finies de p partitions une distance δ telle que

$$\delta [(P_1, P_2, \ldots, P_p), (P'_1, P'_2, \ldots, P'_p)] = \\ = \frac{1}{p} \sum_{i=1}^{p} d(P_i, P'_i) \quad (1)$$

Si $(P_1, P_2, \ldots, P_p)$ est la famille donnée de partitions, on préférera d'autant mieux une partition P que la distance selon δ de $(P, P, \ldots, P)$ à $(P_1, P_2, \ldots, P_p)$ est plus petite et on appellera *partitions centrales* celles qui minimisent

$$\frac{1}{p} \sum_{i=1}^{p} d(P_i, P) . \quad (2)$$

3.2 — *Expression simplifiée du critère défini.* Etant donné une partition P, introduisons les $n \times n$ variables logiques

$$\begin{cases} \varpi_{xy} = 1 \text{ si } x \text{ et } y \text{ sont dans la même classe} \\ \quad = 0 \text{ sinon} . \end{cases}$$

La fonction $(x, y) \xrightarrow{\varpi} \varpi_{xy}$ est la fonction caractéristique du graphe dans $E \times E$ de la relation d'équivalence associée à Π. ϖ est une application de $E \times E$ sur l'ensemble à deux éléments $\{0, 1\}$. C'est donc un élément de $\{0, 1\}^{E \times E}$.

L'ensemble $\mathcal{P}$ des partitions sur E est la partie de ce cube pour laquelle :

$$P \in \mathcal{P} \Longleftrightarrow \begin{cases} \varpi_{xy} = \varpi_{yx} \\ \\ \varpi_{xy} = 1 \text{ et } \varpi_{yz} = 1 \Longrightarrow \varpi_{xz} = 1 . \end{cases}$$

Plongeant le cube des relations, $\{0, 1\}^{E \times E}$, dans $R^{E \times E}$, lequel est isomorphe à $R^{n \times n}$, la métrique euclidienne de $R^{n \times n}$ définit une distance D sur l'ensemble des relations binaires sur E, de la façon suivante

$$D^2 = (\rho - \rho')^2 = \sum_{(x,y) \in E \times E} (\rho_{xy} - \rho'_{xy})^2 \; ; \quad (1)$$

dans cette formule ρ et ρ' sont deux relations de sorte que ρ_{xy} et ρ'_{xy} prennent leurs valeurs dans $\{0, 1\}$. D^2 est égal exactement au cardinal de la différence symétrique des graphes de ρ et de ρ' dans $E \times E$. Dans le cas où ρ et ρ' définissent des partitions, $D^2 = d(\rho, \rho')$; c'est-à-dire encore le nombre de couples, (x, y), qui sont classés ensemble par une seule des deux partitions. Dans ces conditions, la formule (2) du paragraphe précédent se met sous la forme

$$\frac{1}{p} \sum_{h=1}^{p} D^2 (P_h, P) . \quad (2)$$

L'intérêt de plonger le cube des relations dans $R^{n\times n}$ tient aux propriétés des barycentres, dit S. Régnier, dont nous allons reprendre les notations.

$s^h = (s^h_{xy})$ représentera le point de $\{0\ ,\ 1\}^{E\times E}$ relatif à la partition P_h, $S = \frac{1}{p}\sum_{h=1}^{p} s^h$, le barycentre des points s^h également pondérés et $(\bar{\omega}_{xy})$ le point associé à P.

L'expression (2) peut être écrite sous la forme :

$$\frac{1}{p}\sum_{h=1}^{p}\sum_{E\times E}(s^h_{xy}-\bar{\omega}_{xy})^2 = \frac{1}{p}\sum_{h=1}^{p}\|s^h-\bar{\omega}\|^2$$

où $\|\bullet\|$ désigne la norme euclidienne définie dans $R^{n\times n}$.

Or

$$\sum_{h=1}^{p}\frac{1}{p}\|s^h-P\|^2 = \|S-P\|^2 + \sum_{1}^{p}\frac{1}{p}\|s^h-S\|^2\ .$$

Par conséquent une partition centrale peut être définie comme réalisant le minimum de $\|P-S\|^2$.

La fonction $\|P-S\|^2$ de P se simplifie si on remarque que $\omega_{xy} = 0$ ou 1 pour tout couple $(x\ ,y)$. En effet le point P reste à une distance fixe du point H, de coordonnées toutes égales à 1/2, lorsqu'il parcourt l'ensemble $\mathfrak{P}$ des partitions ou même l'ensemble de toutes les relations binaires. En effet

$$\|P-H\|^2 = \sum_{(x,y)\in E\times E}\left(\omega_{xy}-\frac{1}{2}\right)^2 = n^2/4\ .$$

En termes géométriques, le cube des relations est inscrit dans la sphère de centre H et de rayon $n^2/4$. Par conséquent

$$\begin{aligned}\|P-S\|^2 &= \|(P-H)-(S-H)\|^2\\ &= \|P-H\|^2 + \|S-H\|^2\\ &\quad - 2<(P-H)\ ,(S-H)> \qquad (4)\end{aligned}$$

où nous avons désigné par $<x\ ,y>$ la forme bilinéaire symétrique associée à la norme euclidienne. (1) s'écrit donc

$$n^2/4 + \|S-H\|^2 - 2<P-H\ ,S-H>\ .$$

Par suite, réaliser le minimum de la norme $\|P-S\|$ revient à maximiser, S étant donné, le produit scalaire $<P-H\ ,\ S-H>$ qui s'exprime sous la forme

$$\sum_{(x,y)\in E\times E}\left(\bar{\omega}_{xy}-\frac{1}{2}\right)\left(s_{xy}-\frac{1}{2}\right)$$

où les s_{xy} sont les coordonnées du point S, c'est-à-dire $s_{xy} = \frac{1}{p}\sum_{h=1}^{p} s^h_{xy}$.

Posant $t_{xy} = s_{xy} - \frac{1}{2}$, une partition centrale est celle qui maximise la forme linéaire

$$L(P) = \sum_{(x,y)\in E\times E} t_{xy}\,\bar{\omega}_{xy}\ . \qquad (5)$$

Relativement à une partition P sur E qui définit les classes $E_1, E_2, \ldots, E_k$, la quantité (5) représente

$$\sum_{h=1}^{k}\left(\sum_{E_h\times E_h} t_{xy}\right) \qquad (6)$$

qu'on peut regarder comme une mesure du graphe de la relation d'équivalence définie par P ; le sommet $(x\ ,y)$ du graphe étant affecté de la masse t_{xy}. $\sum_{E_h\times E_h} t_{xy}$ est une sorte de compacité de la classe E_h et $L(P)$ est la somme des compacités des différentes classes. On peut aisément montrer qu'une condition nécessaire et suffisante pour que la partition la plus fine soit optimale au sens de $L(P)$ est que $s_{xy} < \frac{1}{2}$ pour tout couple $(x\ ,y)$ de $E\times E$. D'autre part si $s_{xy} > \frac{1}{2}$ pour tout couple $(x\ ,y)$, la partition la moins fine est optimale ; il serait faux, par contre, de croire qu'il est nécessaire, pour que la partition la moins fine soit centrale, que $s_{xy} > \frac{1}{2}$ pour tout couple $(x\ ,y)$.

Enfin pour s_{xy} constant, égal à 1/2, toute partition est centrale.

Exemple de calcul de (6) : soit à comparer les deux classifications :

$$\{\{a\,,d\,,f\}\quad,\{b\,,c\,,e\}\} \qquad \text{(I)}$$

et

$$\{\{a\,,d\,,e\,,f\},\{b\,,c\}\} \qquad \text{(II)}$$

pour le tableau d'incidence du paragraphe II précédent. Etablissons pour cela le tableau des t_{xy} à partir du tableau des $(s\,,u\,,v)$ défini au chapitre 1 (§ II.2), on a

$$s_{xy} = \frac{s+t}{T} \quad \text{et} \quad t_{xy} = s_{xy} - \frac{1}{2}\,.$$

	a	d	f	e	b	c
a	0,5	$\frac{1}{6}$	$\frac{1}{9}$	$-\frac{1}{6}$	$-\frac{1}{9}$	$-\frac{1}{9}$
d	$\frac{1}{6}$	0,5	$\frac{1}{9}$	$\frac{1}{18}$	0	$-\frac{1}{9}$
f	$\frac{1}{9}$	$\frac{1}{9}$	0,5	0	$-\frac{1}{6}$	$-\frac{7}{18}$
e	$-\frac{1}{6}$	$\frac{1}{18}$	0	0,5	$\frac{1}{9}$	0
b	$-\frac{1}{9}$	0	$-\frac{1}{6}$	$\frac{1}{9}$	0,5	$\frac{1}{6}$
c	$-\frac{1}{9}$	$-\frac{1}{9}$	$-\frac{7}{18}$	0	$\frac{1}{6}$	0,5

Pour (I) la valeur du critère est $\frac{41}{18}+\frac{37}{18}=\frac{13}{3}$ et pour la partition (II) $\frac{46}{18}+\frac{24}{18}=\frac{35}{9}$.

La partition (I) est donc meilleure que la partition (II) du point de vue de ce critère.

3.3 — *Bonnes métriques de* $R^{n\times n}$

Si pour la métrique euclidienne, la recherche de la partition centrale conduit à maximiser une forme linéaire, cela provient du fait que dans $R^{n\times n}$ les différents points du cube des relations, $\{0\,,1\}^{n\times n}$, sont tous équidistants d'un même point H. On peut chercher à caractériser les formes linéaires symétriques non dégénérées définissant des normes ayant cette propriété ; c'est l'objet de la proposition suivante.

Proposition. Pour que la norme associée à une forme bilinéaire symétrique non dégénérée $B\,(z\,,t)$ rende équidistants d'un même point G de $R^{n\times n}$ tous les points de $\{0\,,1\}^{n\times n}$, il est nécessaire et suffisant que la base canonique de $R^{n\times n}$ soit orthogonale.

La condition suffisante est évidente, on montrerait aisément que si la base canonique est orthogonale pour B, tous les points de $\{0\,,1\}^{n\times n}$ sont équidistants du point H de coordonnées toutes égales à 1/2.

Réciproquement, si tous les points du cube $\mathfrak{C}$ des relations sont équidistants d'un même point G, vérifions d'abord que ce point G ne peut être que H. En effet $Q(x)=B(x\,,x)$ désignant la métrique associée à la forme B et X un point courant de $\mathfrak{C}$, on a $Q(X-H)=Q(X'-H)$ où $X'=I-X$ et I le point dont toutes les coordonnées sont égales à 1. Par ailleurs, comme $Q(X-G)=Q(X'-G)$, on a, en développant $Q(X-X')$ de deux façons différentes :

$$\begin{aligned} Q(X-X') &= Q(X-H)+Q(X'-H)-2B(X-H\,,X'-H)\\ &= Q(X-G)+Q(X'-G)-2B(X-G\,,X'-G),\\ B(G-H\,,G-H) &= Q(G-H)=0. \end{aligned}$$

Donc $G=H$ puisque la forme B n'est pas dégénérée.

D'autre part si $Q(X-H)=\text{cte}=K^2$ pour $X\in\mathfrak{C}$, $Q(Y)=4K^2$ pour tout point Y dont les coordonnées sont égales à 1 ou à -1. Il en résulte que Q ne contient pas de termes rectangle dans son développement par rapport à la base canonique, car, s'il n'en était pas ainsi, $Q(Y)$ changerait de valeur en remplaçant Y par $Y'=-Y$. $Q(X)$ est donc de la forme $\Sigma\, q_{xy}\,\xi_{xy}^2$.

On peut noter que si X et Z sont deux relations, $Q(X-Z)$ est encore la mesure de la différence symétrique des graphes de X et de Z pour la mesure q_{xy} définie sur $E\times E$ par les coefficients de la forme Q par rapport à la base canonique.

4 — Objets équivalents

Relativement à la description des objets de E définie par l'application a [cf. § 1 ci-dessus], il peut exister des couples d'objets x et y "identiques à tous les points de vue" qui ont donc même représentation $a(i)$ dans l'ensemble

$$A_1 \times A_2 \times \ldots \times A_p = \mathcal{A} \quad .$$

$$(\forall_h) \quad a_h(x) = a_h(y)$$

On dira qu'ils sont équivalents.

On peut naturellement s'attendre à ce que deux tels objets soient classés ensemble dans les partitions centrales.

En effet, si deux objets x et y de même représentation ($a(x) = a(y)$) étaient séparés pour une partition $P : E_1, E_2, \ldots, E_k$ de E, (par exemple $x \in E_j$ et $y \in E_h$), l'une des deux opérations : transférer l'objet x dans E_k ou bien transférer l'objet y dans E_j, augmente strictement la forme linéaire

$$L(P) = \sum_{h=1}^{k} \left(\sum_{(x,y) \in E_h \times E_h} t_{xy} \right)$$

considérée ci-dessus.

Pour s'en rendre compte, le premier transfert provoque une variation de $L(P)$ qui vaut

$$\Delta_1 [L(P)] = \frac{1}{2} + \sum_{l \in E_h} t_{xl} - \sum_{l \in E_j} t_{xl} \quad ,$$

et le second transfert une variation de $L(P)$ qui vaut

$$\Delta_2 [L(P)] = \frac{1}{2} + \sum_{l \in E_j} t_{xl} - \sum_{l \in E_h} t_{xl} \quad .$$

Or la somme $\Delta_1 [L(P)] + \Delta_2 [L(P)]$ vaut 1, donc l'un des deux accroissements Δ_1 ou Δ_2 est strictement positif.

Puisque deux objets équivalents sont destinés à être dans la même classe de toute partition centrale, il semble naturel de les grouper dès le début et de les traiter comme un seul objet. Le problème initial, portant sur les objets de E, est ainsi remplacé par le "problème aux images" portant sur les m images ($m < n$)

$$v = a(x) \in a(E) \subset A_1 \times A_2 \times \ldots \times A_p = \mathcal{A} \quad .$$

Si P est une partition centrale du problème aux objets, elle définit une partition P' des images, puisque deux objets ayant la même image sont dans une même classe de P et l'on a

$$\varpi_{xy} = \varpi'_{a(y)\,a(y)} \quad .$$

Posons $a(x) = v$ et $a(y) = w$; si n_v et n_w sont les nombres d'objets représentés respectivement par v et w, on peut les regrouper dans l'expression $L(P)$

$$L(P) = \sum_{(x,y)} (\varpi_{xy} - s_{xy})^2 = \sum n_v n_w (\varpi'_{vw} - s'_{vw})^2 \qquad (1)$$

où l'on a posé $s'_{vw} = \frac{1}{p} \sum_{1}^{p} s^h_{vw}$, la dernière somme est étendue à tous les couples d'images (v, w) où v et w appartiennent à $a(E)$. En vertu de ce qui précède et de l'identité (1) on a la

4.1 — *Proposition.* Le problème aux m images conduit aux même partitions centrales que le problème aux objets, à condition d'utiliser sur $R^{m \times m}$ la métrique associée à la forme quadratique

$$Q(P') = \Sigma\, n_v\, n_w\, (\varpi'_{vw})^2 \quad .$$

Cette nouvelle métrique utilisée sur $R^{m \times m}$ présente les mêmes propriétés algébriques que celles du problème initial, notamment celle de barycentre exprimée au paragraphe 3.2 et celle définie par la proposition du paragraphe 3.3. Il faut toutefois noter que, si le carré de la distance entre deux relations X et Y était le cardinal de la différence symétrique de leurs graphes dans $E \times E$, dans le cas présent si X et Y sont deux relations sur

$$\mathcal{A} = A_1 \times A_2 \times \ldots A_p \quad ,$$

$Q(X - Y)$ représente la somme des produits $n_v n_w$ étendue aux couples (v, w) qui vérifient l'une des relations et pas l'autre. C'est donc une mesure positive (à valeurs entières) de la différence symétrique des graphes dans $\mathcal{A} \times \mathcal{A}$.

V — *CRITERES DEFINIS A PARTIR DE LA DONNEE D'UN INDICE DE SIMILARITE*

A — RECHERCHE D'UNE PARTITION

1 — Introduction

Au paragraphe précédent nous avons vu s'introduire subrepticement, relativement à un couple d'objets (x, y) de $E \times E$, le nombre réel positif

$$s_{xy} = \frac{1}{p} \sum_{h=1}^{p} s_{xy}^{h}$$

qui représente la proportion des caractères ayant la même modalité sur les objets x et y ; dans le cas où chaque caractère est une attribution, avec les notations du chapitre 2, le nombre s_{xy} peut être écrit sous la forme $(s + t)/T$. Dans ce paragraphe, nous supposons, plus généralement, défini sur E un indice de similarité qui, rappelons-le, est une fonction réelle positive sur $E \times E$, $S(x, y)$, censée mesurer les ressemblances entre objets. S est donc un élément de $R_{+}^{E \times E}$, qui est isomorphe à $R^{n \times n}$.

On consigne en général la valeur d'un indice de similarité pour E dans un tableau carré, $n \times n$, dont les lignes d'une part et les colonnes d'autre part sont indexées, de la même façon, par les objets de E ; la case relative à la ligne x et à la colonne y porte le nombre $S(x, y)$.

2 — Critères définis pour la recherche d'une partition

2.1 — Sans nuire à la généralité, on peut presque toujours se ramener au cas où $S(x, y)$ prend ses valeurs dans l'intervalle $[0, 1]$ et où $S(x, x) = 1$ pour tout x de E. C'est ce que suppose d'ailleurs S. Régnier pour adapter son critère au choix d'un indice de similarité quelconque.

S désignant le point de $R^{n \times n}$ représentant l'indice de similarité et P le point de $\{0, 1\}^{E \times E}$ associé à une partition, moyennant le plongement du cube des relations dans $R^{n \times n}$, on préférera d'autant mieux une partition P que $\|P - S\|$ est plus petite, où $\|\bullet\|$ est la norme euclidienne de $R^{n \times n}$.

Les résultats établis au paragraphe précédent se transposent exactement ici. Notons toutefois que, compte tenu du caractère crucial de la valeur 1/2 pour l'indice de similarité [cf. § 3.2], relativement à la perception de la ressemblance entre les objets de E qu'on s'en fait, il importe que $S(x, y) = 1/2$ corresponde au cas où x et y sont "moyennement" ressemblants.

2.2 — Dans la mesure où, d'une part, le nombre de classes peut être naturellement fixé à priori et où, d'autre part, on peut admettre de plonger l'espace des descriptions (ici $\mathcal{A} = A_1 \times A_2 \times \ldots \times A_p$) dans un espace préhilbertien (espace vectoriel muni d'un produit scalaire), nous allons signaler un critère présenté par Vo Khac Khoan et Nghiem Phong Tuan dans un récent article [36].

On associe à une classe E_i son vecteur moyen

$$E_i \longrightarrow \vec{g}(E_i) = \frac{1}{|E_i|} \sum_{\vec{x} \in E_i} \vec{x} \ .$$

On définit la dispersion interne à E_i par

$$\Delta(E_i) = \sum_{\vec{x} \in E_i} \|\vec{x} - \vec{g}(E_i)\|^2 \ .$$

Relativement aux différentes classes, on définira la dispersion de mélange par

$$\Delta(E_1, E_2, \ldots, E_k) = \sum_{i=1}^{k} \|\vec{g}(E_i) - \vec{G}\|^2$$

où $\vec{G}$ désigne le vecteur moyen de la population.

Le critère consiste à considérer une partition d'autant préférable que la dispersion de mélange des différentes classes qui lui sont relatives est plus grande. Il est rassurant de savoir que maximiser la dispersion de mélange de $(E_1, E_2, \ldots, E_k)$ revient à minimiser la somme des dispersions internes des différentes classes ; on s'en rend compte par un calcul qui, comme celui du paragraphe IV.3.2., utilise les propriétés des barycentres.

B — RECHERCHE D'UNE CHAINE DE PARTITIONS

1 — Représentation d'une chaine de partitions

La donnée étant une métrique ou pseudo-métrique $D'(x, y) = 1 - S(x, y)$ (où $S(x, y)$ est un indice de similarité à valeurs dans $[0, 1]$), on représentera, naturellement, une chaîne de partitions par la donnée sur E d'une distance ultramétrique en vertu de la proposition 10 (Chap. 0 § 4).

Nous verrons au paragraphe suivant la forme que nous avons donnée à cette représentation ; en même temps qu'à nous-mêmes, cette représentation s'est imposée à plusieurs auteurs, notamment J.P. Benzécri et M. Roux, C.J. Jardine, N. Jardine et R. Sibson, S.C. Johnson.

Se référant à la donnée d'une matrice de distances [cf. § A.1], M. Eytan caractérise celles qui correspondent à une distance ultramétrique de la façon suivante ;

Proposition. Pour qu'une matrice $(n \times n)$ de nombres réels positifs, symétrique par rapport à la diagonale, soit une matrice de distances associée à une distance ultramétrique sur E, il faut et il suffit qu'on puisse l'amener par permutation des lignes et des colonnes de même nom à la forme suivante :

a) au-delà du terme diagonal qui vaut 0, les éléments d'une même ligne vont en croissant (au sens large) ;

b) Pour tout k si $d(k, k+1) \neq d(k, k+2)$ alors $d(k, j) = d(k+1, j)$ pour tout $j > k+1$; si par contre

$$d(k, k+1) = d(k, k+2) = \ldots = d(k, k+l+1)$$

alors $d(k+1, j) \leqslant d(k, j)$ pour $k+1 < j \leqslant k+l$ et $d(k+1, j) = d(k, j)$ pour $j > k+l$.

En effet, soit (E, d) un espace ultramétrique. Désignons par $x_i^{(j)}$ les éléments à distance i de x_0, c'est-à-dire

$$(\forall i, j)\ d(x_0, x_i^{(j)}) = i \iff (\forall i, j)\ x_i^{(j)} \in S(x_0; i)\ .$$

Ordonnons maintenant E suivant i, et pour des x de même indice i, suivant un ordre quelconque concrétisé par le deuxième indice j. Autrement dit $y = x_i^{(j)}$ si et seulement si $d(x_0, y) = i$ i.e.

$$y \in C(x_0; i) - C(x_0; i-1)\ .$$

Posons $K(x_0; i) = C(x_0; i) - C(x_0; i-1)$ (que nous appellerons "couronne circulaire de rang i, de centre x_0"). Les $K(x_0; i)$ forment, lorsque i décrit N (en fait lorsque i décrit un segment fini de N), une partition de E. D'après la Proposition 2 § IV

$$C(x_0; i_0) = C(x_{i_0}^{(j)}; i_0)$$

puisque $x_{i_0}^{(j)} \in C_{(x_0; i_0)}$. De plus, d'après le Corollaire 2.1, les $C(x_{i_0}^{(j)}; i)$, pour $i > i_0$ contiennent $C(x_{i_0}^{(j)}; i_0)$ donc $C(x_0; i) = C(x_{i_0}^{(j_0)}; i)$ pour tout $i > i_0$.

On en déduit donc que $K(x_0; i) = K(x_{i_0}^{(j_0)}; i)$ pour $i > i_0$. Ainsi :

$$\begin{aligned} \forall i > i_0\ , x_i^{(j)} \in K(x_0; i) &\iff x_i^{(j)} \in K(x_{i_0}^{(j_0)}; i) \\ &\iff d(x_{i_0}^{(j_0)}; x_i^{(j)}) = i \\ &= d(x_0, x_i^{(j)})\ , \end{aligned}$$

ce qui démontre la première partie de b), en prenant $i = i_0 + 1$, j_0 et j tels que $x_i^{(j)}$ soit successeur immédiat de $x_{i_0}^{(j_0)}$.

Si maintenant

$$i = i_0, d(x_{i_0}^{(j_0)}, x_{i_0}^{(j)}) \leqslant \text{Max}\,(d(x_{i_0}^{(j_0)}, x_0)\ , d(x_0, x_{i_0}^{(j)})) = i_0$$

soit $d(x_{i_0}^{(j_0)}, x_{i_0}^{(j)}) \leqslant i_0$, ce qui démontre la deuxième partie de b), en prenant $j = j_0 + 1$. Quant à a), elle est démontrée :

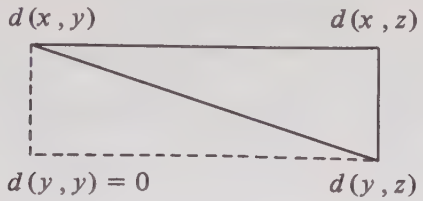
pour la première ligne par la façon dont on a ordonné E ;

pour les autres lignes en appliquant b) et en modifiant les j pour les $x_{i_0}^{(j)}$ éventuellement.

Inversement, soit d une matrice symétrique vérifiant a), b). La condition de symétrie assure tous les axiomes sauf (UM).

Vérifions que b) et c) l'entraînent. Pour cela considérons les éléments x et z de E fixes, et vérifions que pour tout $y \in E$, (UM) est vrai. Ordonnons auparavant E dans l'ordre des "marges" de d. A un automorphisme près de cet ordre, l'élément y est soit intérieur à $]x, z[= \{y : x < y < z\}$ soit extérieur à cet intervalle.

1^{er} *cas* : $x < y < z$. Considérons le triangle



(I) si $d(x, y) \neq d(x, z)$ i.e. par a), si

$$d(x, y) < d(x, z)$$

alors $d(y, z) = d(x, z)$ par b). On a bien

$$d(x, z) = \text{Max}\,(d(x, y), d(y, z)) = d(y, z)$$

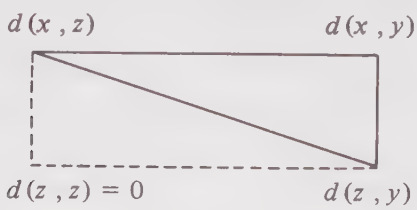
(II) si $d(x, y) = d(x, z)$, alors pour b)

$$d(y, z) \leqslant d(x, z)$$

et encore

$$d(x, z) \leqslant \max\,[d(x, y), d(y, z)) = d(x, y)\,.$$

$2^{ème}$ *cas* : $x < z < y$. Considérons le triangle



Par a), on a $d(x, z) \leqslant d(x, y)$ et par b)

$$d(z, y) \leqslant d(x, y)$$

Donc $d(x, z) \leqslant \max\,(d(x, y), d(y, z)) = d(x, y)$.

A titre d'illustration on se reportera à la matrice des distances figurées à la fin du chapitre 0.

2 — Définition d'un critère

On peut vouloir, profitant de la caractérisation qui vient d'être donnée d'une matrice de distances ultramétrique, chercher celles des matrices ultramétriques, Δ, qui minimisent la distance, au sens de la métrique euclidienne dans $R^{n \times n}$, entre Δ et la matrice D des distances (ou pseudo-distances) donnée sur E :

$$\|\Delta - D\|^2 \text{ minimum}\,.$$

Les matrices $D = (d_{ij})$ et $\Delta = (\delta_{ij})$ étant symétriques, d'autre part ayant $d_{ii} = \delta_{ii} = 0$ pour tout i, le problème revient à minimiser

$$\sum_{i<j} (\delta_{ij} - d_{ij})^2$$

sous les contraintes

$$\delta_{ij} \leqslant \max\,(\delta_{ik}, \delta_{kj})$$

pour tout i, j et k de E.

Le problème qui vient d'être posé est extrêmement complexe ; il est de même très difficile de concevoir un algorithme, qui opérerait par permutation des lignes et colonnes de D respectivement associées, aboutissant à un optimum local cohérent avec le critère défini ; comme ce sera le cas pour la recherche d'une partition [cf. Ch. 3 § II "Algorithme des transferts"].

Dans une récente publication, [11], C.J. Jardine, N. Jardine et R. Sibson, définissant sur l'ensemble des métriques sur E la relation de dominance (i.e. d'ordre) suivante

$$d^{(1)} \geqslant d^{(2)} \Longleftrightarrow d^{(1)}(x, y) \geqslant d^{(2)}(x, y)$$

(pour tout x et y de E) qu'on peut lire $d^{(1)}$ domine $d^{(2)}$, proposent de préférer la distance ultramétrique δ qui réalise $\sup\{\delta \mid \delta \in \Delta \text{ et } \delta \leqslant d\}$ où Δ est l'ensemble des distances ultramétriques et d la métrique définie sur E.

Il suffit de remarquer que la chaîne de partitions définie par δ est indépendante d'une transformation strictement croissante des $d(x, y)$ ou des $\delta(x, y)$ pour comprendre pourquoi, indépendamment de ce travail, nous avions exprimé ce même critère, (ainsi d'ailleurs que d'autres) en termes de proximité entre préordonnances sur E.

Ne tardons donc plus à nous placer dans le cadre assez général et relativement stable [cf. Chap. 1 § III] où la donnée est une préordonnance sur E pour étudier différents critères et les comparer. Bien qu'il soit assez général ce cadre est suffisant pour obtenir des résultats féconds.

VI — *CRITERES DEFINIS A PARTIR D'UNE PREORDONNANCE SUR E*

Origine de la donnée.

La donnée de base est ici un préordre total sur l'ensemble F des paires d'objets distincts de E :

$$F = \{\{x, y\}; x \in E, y \in E \text{ et } x \neq y\}.$$

Une paire p précède une paire q ($p \leqslant q$), si les deux objets composant p se ressemblent davantage que ceux composant q ; l'origine de cette relation d'ordre a été explicitée dans le chapitre 1 auquel on se reportera. ω désigne le graphe dans $F \times F$ du préordre.

A — RECHERCHE D'UNE PARTITION

1 — Représentation d'une partition

La donnée d'une partition sur E peut être caractérisée par la donnée sur F d'un préordre à deux classes : R (resp. S) ensemble des paires réunies (resp. séparées) où R satisfait la condition.

$$\{x, y\} \in R \text{ et } \{y, z\} \in R \Longrightarrow \{x, z\} \in R$$

et où toute paire réunie précède toute paire séparée.

2 — Définition d'un critère

2.1 — *Critère de J.P. Benzécri*

J.P. Benzécri avait proposé de préférer d'autant plus une partition P que le cardinal de l'intersection, dans $F \times F$, de ω et de $R \times S$ est plus grand. De la sorte, le problème de la recherche d'une classification revient à maximiser $|\omega \cap R \times S|$ où R et S sont relatifs à une partition P.

2.2 — *Critère de W.F. de la Véga*

W.F. de la Véga proposa, pour la recherche d'une classification, de maximiser le cardinal du complémentaire dans $F \times F$, de la différence symétrique entre ω et $R \times S$, soit $|F \times F - \omega \Delta R \times S|$.

En tenant compte des relations :

$$|\omega \cup R \times S| = |\omega| + |R \times S| - |\omega \cap R \times S|$$
$$= |\omega \Delta R \times S| + |\omega \cap R \times S|,$$

on obtient :

$$|F \times F - \omega \Delta R \times S| = |F|^2 - |\omega| + 2\left(|\omega \cap R \times S| - \frac{1}{2}|R| \times |S|\right).$$

Relativement au problème d'optimisation, il s'agit donc ici de rendre maximum

$$|\omega \cap R \times S| - \frac{1}{2}|R| \times |S|$$

où ω est donné et où R et S sont relatifs à une même partition P, objet de notre recherche.

Introduisant les fonctions indicatrices I_ω et $I_{R \times S}$ des sous-ensembles ω et $R \times S$ de $F \times F$, le premier critère se met sous la forme :

$$\sum_{F \times F} I_{R \times S}(p, q)\, I_\omega(p, q) \qquad (1)$$

et le second :

$$\sum_{F \times F} I_{R \times S}(p, q)\left[I_\omega(p, q) - \frac{1}{2} \right] \qquad (2)$$

Les expressions (1) et (2) sont des formes linéaires définies sur l'espace $\{0,1\}^{F\times F}$; (2) n'est pas sans nous rappeler le produit scalaire (5) du paragraphe IV.3.2.

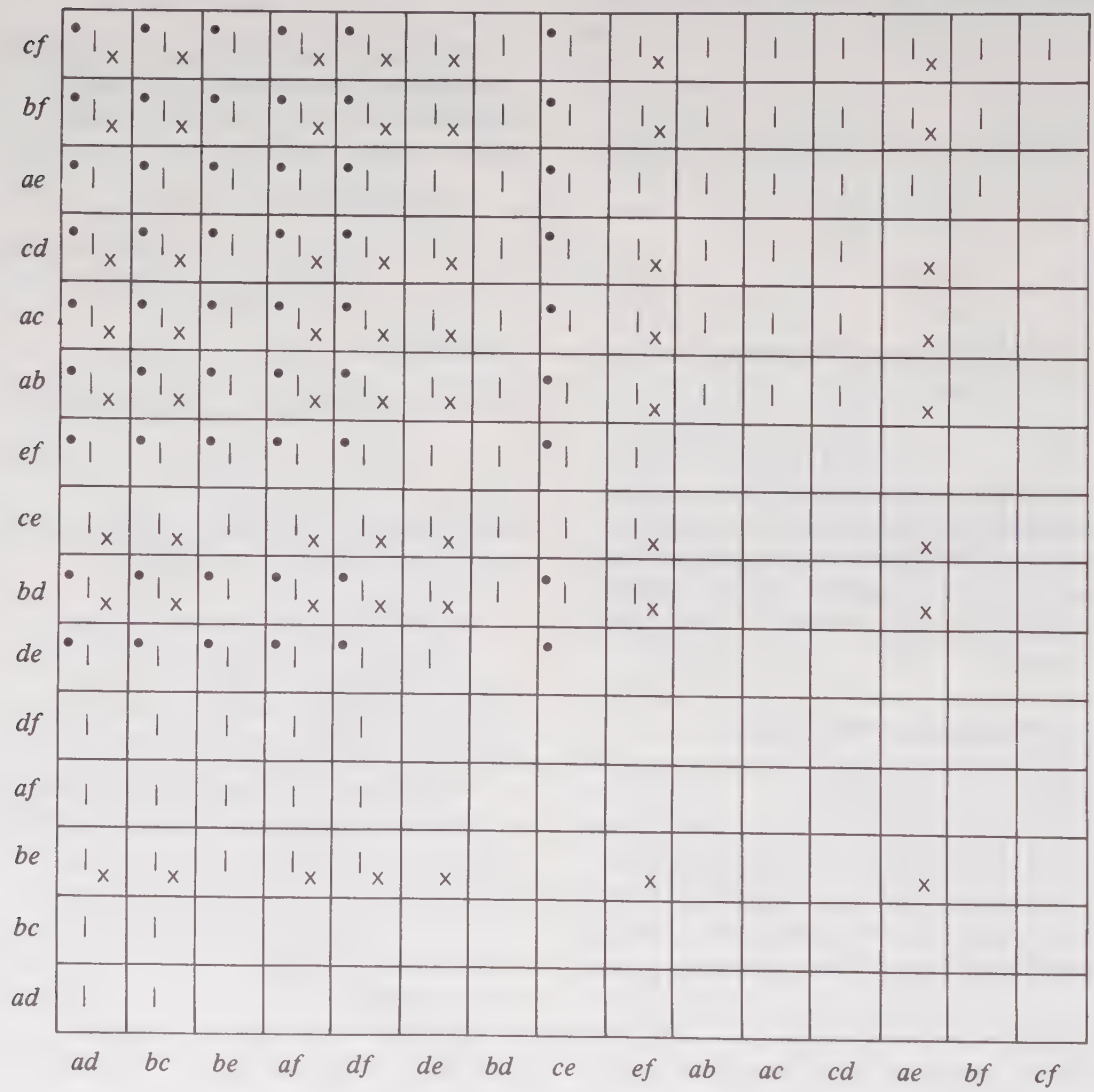
Exemple : au vu de la préordonnance $\overline{\omega}$

$$\{a,d\}=\{b,c\}<\{b,e\}=\{a,f\}=\{d,f\}<\{d,e\}<$$
$$<\{b,d\}=\{c,e\}=\{e,f\}<\{a,b\}=\{a,c\}=$$
$$=\{c,d\}<\{a,e\}=\{b,f\}<\{c,f\}\quad,$$

établie á partir du tableau d'incidence des données (Ch. 1, § II.2) sur la base de l'indice $(u+v)$, nous allons comparer les deux partitions

$$\{\{a,d,f\}\quad,\{b,c,e\}\} \qquad \text{(I)}$$
$$\{\{a,d,e,f\},\{b,c\}\} \qquad \text{(II)}$$

du point de vue de chacun des deux critères précédents.



Les traits verticaux représentent le graphe de $\overline{\omega}$ dans $F \times F$, les points, le rectangle $R \times S$ associé à la partition (I) et les croix, le rectangle $R \times S$ associé à la partition (II).

Pour la partition (I), on a

$$|\overline{\omega} \cap R \times S| = 53 ,$$

d'où

$$|\overline{\omega} \cap R \times S| - \frac{1}{2}|R| \times |S| = 53 - \frac{1}{2} 6 \times 9 = 25 .$$

Pour la partition (II), on a

$$|\overline{\omega} \cap R \times S| = 48 ,$$

d'où

$$|\overline{\omega} \cap R \times S| - \frac{1}{2}|R| \times |S| = 48 - \frac{1}{2} 7 \times 8 = 20 .$$

La partition (I) est préférable à la partition (II) de chacun des deux points de vue.

Avant d'aller plus loin dans la proposition d'autres critères nous allons établir certaines propriétés statistiques liées à (1) ou à (2), ces propriétés nous permettront en particulier de comparer les deux critères.

3 — Comparaison des deux critères

3.1 — *Préliminaires.* Sur quelques problèmes particuliers d'énumération liés aux partitions.

a) Rappelons que $\mathfrak{F}$ désigne l'ensemble des partitions aux classes étiquetées de type $(n_1, n_2, \ldots, n_k)$, donné, pour lesquelles $|f^{-1}(i)| = n_i$ et que Λ représente celui des partitions aux classes non étiquetées de type $(n_1, n_2, \ldots, n_k)$. [cf. Ch. 0 § III].

Rappelons aussi que pour un élément P de Λ

$$|R| = \sum_{1 \leqslant i \leqslant k} n_i (n_i - 1)/2$$

et

$$|S| = \sum_{i<j} n_i n_j ,$$

d'autre part $|R| + |S| = |F| = n(n-1)/2$.

b) *Lemmes*

b.1 — La proportion d'éléments de $\mathfrak{F}$ pour lesquels une paire d'objets donnée $\{x , y\}$ est contenue dans une même classe est $\sum_{i=1}^{k} n_i (n_i - 1)/n(n-1)$. Il s'agit de déterminer le cardinal du sous-ensemble de $\mathfrak{F}$: $\{f \in \mathfrak{F} | f(x) = f(y)\}$. Or

$$\{f \in \mathfrak{F} | f(x) = f(y)\} = \sum_{i=1}^{k} \{f \in \mathfrak{F} | f(x) = f(y) = i\} ;$$

somme ensembliste et

$$|\{f \in \mathfrak{F} | f(x) = f(y)\}| = \sum_{i=1}^{k} |\{f \in \mathfrak{F} | f(x) = f(y) = i\}|$$

Le sous-ensemble $\{f(x) = f(y) = i\}$ peut être mis en correspondance biunivoque avec l'ensemble des "partitions aux classes étiquetées" de $E - \{x , y\}$, telles que les effectifs des différentes classes d'une même partition, sous la réserve $n_i \geqslant 2$, soient respectivement $n_1, n_2, \ldots, n_{(i-1)}, n_i - 2, n_{(i+1)}, \ldots, n_k$. Donc

$$|\{f(x) = f(y) = i\}| = (n-2)!/n_1! \, n_2! \ldots (n_i - 2)! \ldots n_k!$$

si $n_i \geqslant 2$ et 0 sinon. D'où

$$|\{f(x) = f(y)\}| = \sum_{\{i, n_i \geqslant 2\}} (n-2)!/n_1! \, n_2! \ldots (n_i - 2)! \ldots n_k!$$

En formant le rapport de ce cardinal sur $|\mathfrak{F}|$, on trouve le résultat annoncé. Nous abrégerons les démonstrations suivantes qui sont analogues à cette dernière.

b.2 — La proportion d'éléments de $|\mathfrak{F}|$, pour lesquels les deux objets d'une paire donnée $\{x , y\}$ appartiennent respectivement à deux classes distinctes, est $2 \sum_{i<j} n_i n_j / n(n-1)$.

$$|\{f \in \mathfrak{F} | f(x) \neq f(y)\}| = \sum_{i \neq j} |\{f \in \mathfrak{F} | f(x) = i , \; f(y) = j\}| ;$$

la somme comporte $k(k-1)$ termes.

L'ensemble $\{f(x) = i\,, f(y) = j \text{ et } i \neq j\}$ peut être mis en correspondance biunivoque avec l'ensemble des "partitions aux classes étiquetées" de $E - \{x\,, y\}$ telles que les effectifs des différentes classes d'une même partition soient respectivement :

$$n_1\,, n_2\,, \ldots\,, n_{(i-1)}\,, n_i - 1\,, n_{i+1}\,, \ldots\,, n_{(j-1)}\,, n_j - 1\,, n_{j+1}\,, \ldots\,, n_k\,.$$

On complètera le calcul et retrouvera le résultat annoncé.

b.3 — Etant donnés trois objets quelconques x, y et z de E, la proportion d'éléments de $\mathfrak{F}$ pour lesquels la paire $\{x\,, y\}$ est réunie et la paire $\{x\,, z\}$ est séparée est $\sum_{i \neq j} n_i (n_i - 1)\, n_j / n\,(n-1)\,(n-2)$;

$$|\{f \in \mathfrak{F} | f(x) = f(y) \neq f(z)\}| = \sum_H |\{f \in \mathfrak{F} | f(x) = f(y) = i \quad \text{et} \quad f(z) = j\}|$$

où $H = \{(i\,, j) \mid i \neq j\,, n_i \geqslant 2\}$

L'ensemble

$$\{f \in \mathfrak{F} | f(x) = f(y) = i\,, f(z) = j\,, i \neq j \quad \text{et} \quad n_i \geqslant 2\}$$

peut être mis en correspondance biunivoque avec l'ensemble des "partitions aux classes étiquetées" de $E - \{x\,, y\,, z\}$, telles que les effectifs des différentes classes, sous la réserve $n_i \geqslant 2$, soient respectivement

$$n_i\,, n_2\,, \ldots\,, n_{(i-1)}\,, n_i - 2\,, n_{i+1}\,, \ldots\,, n_{(j-1)}\,, n_j - 1\,, n_{j+1}\,, \ldots\,, n_k\,.$$

Cette remarque permet d'exprimer le terme courant de la somme ci-dessus et d'établir donc le résultat.

b.4 — Etant donnés quatre objets x, y, z et t de E, la proportion d'éléments de $\mathfrak{F}$, pour lesquels la paire $\{x\,, y\}$ est réunie et la paire $\{z\,, t\}$ séparée est

$$\left[\sum_J n_i (n_i - 1)\, n_j\, n_l + 2 \sum_K n_i (n_i - 1)(n_i - 2)\, n_j\right] \Big/ n\,(n-1)\,(n-2)\,(n-3)$$

où $J = \{(i\,, j\,, l) | i \neq j\,, j \neq l \text{ et } l \neq i\,, n_i \geqslant 2\}$ et $K = \{(i\,, j) | i \neq j\,, n_i \geqslant 3\}$. La première somme comporte $k\,(k-1)\,(k-2)$ termes au plus et la seconde $k\,(k-1)$ termes au plus.

Pour qu'un élément f de $\mathfrak{F}$ réalise $f(x) = f(y)$ et $f(z) \neq f(t)$, il est nécessaire et suffisant d'avoir

1/ $f(x) = f(y) = i\,, f(z) = j$ et $f(t) = l$ pour un triplet $(i\,, j\,, l)$ tel que $i \neq j\,, j \neq l\,, l \neq i$ et $n_i \geqslant 2$ ou bien

2/ a — $f(x) = f(y) = f(z) = i$ et $f(t) = j$ ou

b — $f(x) = f(y) = f(t) = i$ et $f(z) = j$

pour un couple $(i\,, j)$ tel que $i \neq j$ et $n_i \geqslant 3$.

Le cardinal du sous-ensemble des éléments de $\mathfrak{F}$ qui satisfont 1 est celui de l'ensemble des partitions étiquetées de $E - \{x\,, y\,, z\,, t\}$, telles que les effectifs des différentes classes, sous la réserve $n_i \geqslant 2$, soient respectivement :

$$n_1\,, n_2\,, \ldots\,, n_{(i-1)}\,, n_i - 2\,, n_{(i+1)}\,, \ldots\,, n_{(j-1)}\,, n_j - 1\,, n_{(j+1)}\,, \ldots\,, n_{(l-1)}\,, n_l - 1\,, n_{l+1}\,, \ldots\,, n_k$$

le cardinal du sous-ensemble des éléments de $\mathfrak{F}$ qui satisfont 2.a., respectivement 2.b., est celui de l'ensemble des partitions étiquetées de $E - \{x\,, y\,, z\,, t\}$, telles que les effectifs des différentes classes sous la réserve $n_i \geqslant 3$, soient respectivement

$$n_1\,, n_2\,, \ldots\,, n_{(i-1)}\,, n_i - 3\,, n_{i+1}\,, \ldots\,, n_{(j-1)}\,, n_j - 1\,, n_{j+1}\,, \ldots\,, n_k\,.$$

En s'aidant de ces remarques on complètera le calcul et retrouvera le résultat annoncé.

b.5 — Les différentes proportions que nous venons de calculer sont les mêmes, si au lieu de les considérer dans $\mathfrak{F}$, on les considérait dans Λ.

En effet, à l'élément P de Λ correspondent, exactement, $k_1!\, k_2! \ldots k_r!$ éléments de $\mathfrak{F}$ en numérotant de $k_i!$ façons les k_i classes de même cardinal ν_i $i = 1\,, 2\,, \ldots\,, r$. [cf. ch. 0. § III]. Supposons avoir calculé une des proportions précédentes (1, 2, 3 ou 4) dans Λ ; on obtient cette même proportion dans $\mathfrak{F}$ en multipliant les deux termes du rapport par

$k_1! \times k_2! \times \ldots \times k_r!$. Car, pour qu'un élément P intervienne dans le décompte de l'un des termes du rapport, il est nécessaire et suffisant que les

$$k_1!\, k_2! \ldots k_r!$$

éléments correspondants de $\mathfrak{F}$ interviennent dans le décompte de ce terme du rapport pour la proportion calculée dans $\mathfrak{F}$.

Désormais nous travaillerons dans Λ.

b.6 — *Remarques importantes*

1/ Par raison de symétrie, les différentes proportions définies sont invariantes quels que soient les objets composant la paire ou le couple de paires.

2/ Il est nécessaire de rappeler ici que

a) le cardinal de l'ensemble G des couples de paires $(p\,,q)$ de la forme $(\{x\,,y\}\,,\{x\,,z\})$, où x, y et z sont mutuellement distincts, est $n(n-1)(n-2)$.

b) le cardinal de l'ensemble H des couples de paires $(p\,,q)$ de la forme $(\{x\,,y\}\,,\{z\,,t\})$, où x, y z et t sont mutuellement distincts est

$$n(n-1)(n-2)(n-3)/4$$

3.2 — *Résultat principal*

Théorème. La moyenne dans Λ de

$$|\omega \cap R \times S| - \frac{1}{2}|R| \times |S|$$

est nulle si $\overline{\omega}$ est un ordre total sur F.

Si P est un élément de Λ auquel correspondent R et S, on a :

$$\frac{1}{2}|R| \times |S| = \frac{1}{4}\Big(\sum_{i=1}^{k} n_i(n_i-1)\Big) \times \Big(\sum_{i<j} n_i\, n_j\Big) \quad (1)$$

Introduisant les fonctions indicatrices I_ω et $I_{R\times S}$ [cf. § 2.2], la moyenne dans Λ de $|\omega \cap R \times S|$ peut être écrite sous la forme

$$\frac{1}{|\Lambda|}\sum_{\omega\in\Pi}\ \sum_{(p,q)\in F\times F} I_\omega(p\,,q)\, I_{R\times S}(p\,,q) \quad (2)$$

Décomposons la somme portée sur $F \times F$ en deux sommes. La première sera portée sur G et la seconde sur H [cf. remarques ci-dessus]

$$\frac{1}{|\Lambda|}\sum_{\Lambda}\sum_{G} I_\omega(p\,,q)\,.\,I_{R\times S}(p\,,q) + \frac{1}{|\Lambda|}\sum_{\Lambda}\sum_{H} I_\omega(p\,,q)\,.\,I_{R\times S}(p\,,q) \quad (3)$$

En invertissant les signes sommes, on peut écrire (3) sous la forme

$$\sum_{G} I_\omega(p\,,q) \times \frac{1}{|\Lambda|} I_{R\times S}(p\,,q) + \sum_{H} I_\omega(p\,,q) \times \frac{1}{|\Lambda|}\sum_{\Lambda} I_{R\times S}(p\,,q) \quad (4)$$

ω étant un ordre total :

$$(p\,,q)\in\omega \iff (q\,,p)\notin\omega\,.$$

D'autre part $(p\,,q)\in G \iff (q\,,p)\in G$ et

$$(p\,,q)\in H \iff (q\,,p)\in H\,.$$

En tenant compte de ces remarques et des lemmes 3 et 4, l'expression (4) peut être écrite sous la forme

$$\frac{1}{2}\,n(n-1)(n-2) \times \sum_{i\neq j} n_i(n_i-1)\,n_j/n(n-1)(n-2) + \frac{1}{8}\,n(n-1)(n-2)(n-3) \times \left[\sum_{j} n_i(n_i-1)\,n_j\,n_l + 2\sum_{i\neq j}(n_i-1)(n_i-2)\,n_j\right] \Big/ n(n-1)(n-2)(n-3) \quad (5)$$

où $J = \{(i\,,j\,,l) \mid i\neq j\,,\ j\neq l\,,\ l\neq i \text{ et } n_i \geqslant 2\}$.

L'expression (5) devient :

$$\frac{1}{2}\left[\sum_{i\neq j} n_i(n_i-1)\,n_j + \frac{1}{4}\left(\sum_{J} n_i(n_i-1)\,n_j\,n_l + 2\sum_{i\neq j} n_i(n_i-1)(n_i-2)\,n_j\right)\right] \quad (6)$$

La dernière somme, de gauche à droite, peut être exprimée sous la forme

$$\sum_{i \neq j} n_i^2 (n_i - 1) n_j - 2 \sum_{i \neq j} n_i (n_i - 1) n_j$$

d'où, en simplifiant, l'expression(6)devient

$$\frac{1}{2}\left[\frac{1}{4} \sum_{i=1}^{k} n_i (n_i - 1) \left(\sum_{I} n_j n_l + 2 n_i \sum_{j \neq i} n_j \right) \right]$$

où $I = \{(j, l) ; j \neq i \text{ et } l \neq i\}$.

On a :

$$\sum_{I} n_j n_l + 2 n_i \sum_{j \neq i} n_j = \sum_{r \neq s} n_r n_s .$$

On obtient donc bien le résultat énoncé.

3.3 — *Comparaison proprement dite*

Au chapitre 0, nous avons étudié dans quelle mesure la donnée de $|\Psi|$, soit de $\sum_i n_i^2$, nous renseignait sur le type de la partition, c'est-à-dire, pour un entier positif m, quel est l'ensemble des types des partitions de E, telles que $\sum_i n_i^2 = m$ [cf. Ch. 0 § III] . Nous désignerons ici par $\Psi(n, m)$ l'ensemble des types de partitions pour lesquels $\sum_i n_i = n$ et $\sum_i n_i^2 = m$, n et m étant donnés.

Rappelons une propriété qui doit ici être présente à notre esprit ; parmi toutes les partitions de E, celles qui réalisent le maximum de $|R \times S|$ sont celles pour lesquelles $|R|$ est le plus voisin de $n(n-1)/4$.

Dans la mesure où $\Psi(n, n(n+1)/2)$ est différent de l'ensemble vide, il existe au moins un type de partition pour lequel $|R|$ est égal à $n(n-1)/4$. Par exemple si n est un carré, la partition de type $((n + \sqrt{n})/2, (n - \sqrt{n})/2)$ réalise $|R| = n(n-1)/4$; $(n + \sqrt{n})/2$ et $(n - \sqrt{n})/2$ sont bien des entiers puisque n et $\sqrt{n}$ sont de même parité.

Nous noterons $\mathcal{R}_r$ l'ensemble des partitions de E pour lesquelles $|R| = r$, où r est un entier donné ; cet ensemble n'est vide que si $\Psi(n, 2r+n)$ l'est. Les propositions suivantes sont des corollaires du théorème 3.2.

Proposition 1. La moyenne dans $\mathcal{R}_r$ de

$$|\omega \cap R \times S| - \frac{1}{2} |R| \times |S|$$

est nulle si ω est un ordre total sur F.

Les ensembles Λ_t, où Λ_t est l'ensemble des partitions de type t appartenant à $\Psi(n, 2r+n)$, forment une partition de $\mathcal{R}_r$:

$$\mathcal{R}_r = \sum_{t \in T} \Lambda_t \text{ (somme d'ensembles)} ,$$

$$T = \Psi(n, 2r+n) .$$

La moyenne de $|\omega \cap R \times S| - \frac{1}{2} |R| \times |S|$ dans chaque ensemble Λ_t est nulle [cf. § 3.2] ; il en est donc de même dans $\mathcal{R}_r$.

Proposition 2. Les deux critères sont équivalents dans $\mathcal{R}_r$.

En effet, dans l'ensemble $\mathcal{R}_r$ les deux critères ne diffèrent que d'une constante dépendant de n et de r.

Proposition 3. ω étant un ordre total sur F et le cardinal, n, de E tel que $\Psi(n, n(n+1)/2)$ différent de l'ensemble vide ; une partition pour laquelle

$$|R| < \frac{n(n-1)}{4} \left(1 - \frac{\sqrt{2}}{2}\right)$$

ou

$$|R| > \frac{n(n-1)}{4} \left(1 + \frac{\sqrt{2}}{2}\right)$$

ne peut être optimale pour le premier critère.

Puisque $\Psi(n, n(n+1)/2)$ n'est pas vide, il existe au moins une partition P pour laquelle $|R| \times |S|$ est égal à $n^2(n-1)^2/2^4$. Si t est un élément quelconque de $\Psi(n, n(n+1)/2)$, d'après le théorème 3.2 , il existe, dans Λ_t, au moins une partition P_t pour laquelle :

$$|\omega \cap R_t \times S_t| - \frac{1}{2} |R_t \times S_t| \geqslant 0 .$$

c'est-à-dire

$$|\omega \cap R_t \times S_t| \geqslant \frac{1}{2^5} n^2 (n-1)^2 . \qquad (1)$$

R_t et S_t sont relatifs à la partition P_t.

P_m désignant une partition optimale pour le premier critère ($|\omega \cap R \times S|$ maximum), R_m et S_m étant relatifs à P_m, on a :

$$|\omega \cap R_m \times S_m| \geqslant |\omega \cap R_t \times S_t| \geqslant \frac{1}{2^5} n^2 (n-1)^2 \quad (2)$$

Mais $|R_m \times S_m| \geqslant |\omega \cap R_m \times S_m|$, donc nécessairement

$$|R_m \times S_m| \geqslant \frac{1}{2^5} n^2 (n-1)^2 \quad (3)$$

En d'autres termes, une partition pour laquelle

$$|R \times S| < \frac{1}{2^5} n^2 (n-1)^2 \quad (4)$$

ne peut être optimale pour le premier critère.

On a

$$|R \times S| = |R| \times |S| = |R| \times (n(n-1)/2 - |R|)$$

En posant $|R| = r$ et $n(n-1)/2 = f$, l'inégalité (4) devient

$$r(f-r) < \frac{f^2}{2^3} \quad (5)$$

laquelle a lieu si et seulement si

$$r < \frac{f}{2}\left(1 - \frac{\sqrt{2}}{2}\right) \text{ ou } r > \frac{f}{2}\left(1 + \frac{\sqrt{2}}{2}\right). \quad (6)$$

La condition $\Psi(n, n(n+1)/2) \neq \emptyset$ n'est pas restrictive, quel que soit l'entier n, un résultat analogue peut être obtenu en remplaçant le second membre de (1) par la valeur maximum que peut atteindre $\frac{1}{2}|R| \times |S|$ dans l'ensemble des partitions de E.

Proposition 4. L'entier n est supposé satisfaire les deux conditions suivantes :

1/ $\Psi(n, n(n+1)/2) \neq \emptyset$

2/ n est divisible par k, $n = h \times k$.

Une partition de E, en k classes de même cardinal h, ne peut être optimale pour le premier critère, si $k > 6$ ou si $k = 1$.

Pour une partition P de E en k classes de même effectif h, on a

$$|R| = kh(h-1)/2 \ . \quad (1)$$

D'après la proposition ci-dessus, une partition pour laquelle $|R| < \frac{n(n-1)}{4}\left(1 - \frac{\sqrt{2}}{2}\right)$ ne peut être optimale pour le premier critère. Donc la partition P ne peut être optimale si

$$\frac{kh(h-1)}{2} < \frac{kh(kh-1)}{4}\left(1 - \frac{\sqrt{2}}{2}\right) \quad (2)$$

c'est-à-dire si

$$k > 4 + 2\sqrt{2} - \frac{5 + 2\sqrt{2}}{h}, \quad (3)$$

inégalité nécessairement satisfaite pour $k > 6$.

D'autre part, toujours d'après la proposition ci-dessus, une partition pour laquelle

$$|R| > \frac{n(n-1)}{4}\left(1 + \frac{\sqrt{2}}{2}\right)$$

ne peut être optimale pour le premier critère. Donc la partition P ne peut être optimale si

$$\frac{kh(h-1)}{2} > \frac{kh(kh-1)}{4}\left(1 + \frac{\sqrt{2}}{2}\right)$$

c'est-à-dire si

$$k < 2(2 - \sqrt{2}) - \frac{3 - 2\sqrt{2}}{h},$$

inégalité nécessairement satisfaite pour $k = 1$ et $n > 1$.

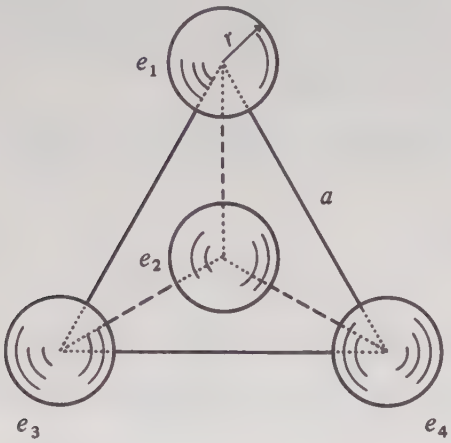
a) Illustrons cette dernière proposition par un exemple. Pour $n = 49$

$$\Psi(n, n(n+1)/2) = \Psi(49, 1225)$$

n'est pas vide, car n est le carré d'un entier et le type $((n + \sqrt{n})/2, (n - \sqrt{n})/2) = (28, 21)$ appartient à $\Psi(49, 1225)$. D'après ce qui précède, une classification d'un ensemble E, de cardinal 49, en 7 classes de même cardinal 7 ne peut être optimale pour le premier critère.

b) Analysons encore l'exemple suivant, emprunté, à W.F. de la Véga : Considérons dans R^3 quatre sphères de même rayon r, e_1, e_2, e_3 et e_4 centrées

respectivement aux sommets d'un tétraèdre régulier d'arête $a > 4r$. E est le solide formé par ces quatre sphères de densité uniforme.



Il est ici plus aisé de travailler dans E^4 plutôt que dans F^2 (remplaçant ainsi F par E^2). R^3 étant muni de la mesure de Lebesgue, la notion de cardinal se généralise par celle de volume. ν désignera le volume commun de l'une quelconque des quatre sphères. L'ordonnance définie sur E est ici naturellement associée à la distance euclidienne.

Considérons les deux partitions suivantes :

$$P_1 = \{e_1, e_2, e_3, e_4\} \text{ et } P_2 = \{e_1 + e_2, e_3 + e_4\},$$

(somme ensembliste).

Pour P_1, l'ensemble des couples réunis, C_r, est défini par

$$C_r = e_1 \times e_1 + e_2 \times e_2 + e_3 \times e_3 + e_4 \times e_4$$

et l'ensemble des couples séparés, C_s, par

$$C_s = e_1 \times e_2 + e_2 \times e_1 + e_1 \times e_3 + e_3 \times e_1 + e_1 \times e_4 + e_4 \times e_1 + e_2 \times e_3 + e_3 \times e_2 + e_2 \times e_4 + e_4 \times e_2 + e_3 \times e_4 + e_4 \times e_3$$

La mesure de $C_r \times C_s$ est donc $4\nu^2 \times 12\nu^2 = 48\nu^4$.

D'autre part deux points d'un couple réuni sont strictement plus proches que deux points d'un couple séparé $(a > 4r)$; il en résulte que $\omega \supset C_r \times C_s$ et la valeur du premier critère est $\underline{48\nu^4}$, tandis que la valeur du second critère est égal à

$$48\nu^4 - 48\nu^4/2 = \underline{24\nu^4} .$$

Pour P_2

$$C_r = (e_1 + e_2) \times (e_1 + e_2) + (e_3 + e_4) \times (e_3 \times e_4)$$

d'où

$$C_r = (e_1^2 + e_1 \times e_2 + e_2 \times e_1 + e_2^2) + (e_3^2 + e_3 \times e_4 + e_4 \times e_3 + e_4^2) .$$

et

$$C_s = (e_1 + e_2) \times (e_3 + e_4) + (e_3 + e_4) \times (e_1 + e_2)$$

La mesure de $C_r \times C_s$ est $8\nu^2 \times 8\nu^2 = 64\nu^4$.

La mesure du sous-ensemble de $C_r \times C_s$,

$$A = \{((x, y), (z, t)) \in C_r \times C_s \mid (x, y) \in e_1^2 + e_2^2 + e_3^2 + e_4^2\}$$

est $4\nu^2 \times 8\nu^2 = 32\nu^4$.

La mesure du sous-ensemble de $C_r \times C_s$,

$$B = \{((x, y), (z, t)) \in C_r \times C_s \mid (x, y) \in e_1 \times e_2 + e_2 \times e_1 + e_3 \times e_4 + e_4 \times e_3\}$$

est $4\nu^2 \times 8\nu^2 = 32\nu^4$.

D'autre part

$$\omega \cap C_r \times C_s = \omega \cap A + \omega \cap B .$$

La mesure de $\omega \cap A$ est $32\nu^4$, $(A \subset \omega)$ est la mesure de $\omega \cap B$ est $16\nu^4$(la moitié de la mesure de B).

Par conséquent la valeur du premier critère est $48\nu^4$ et celle du second critère $48\nu^4 - 32\nu^4 = 16\nu^4$.

Ainsi le premier critère ne discrimine pas entre les deux partitions alors que le second avantage la partition P_1 qui paraît ici la plus naturelle.

3.4 — *Conclusion*

Résumons-nous ; nous avons montré que dans l'ensemble des partitions de type donné, la moyenne

de $|\omega \cap R \times S| - \frac{1}{2}|R| \times |S|$ est nulle, donc indépendante du type considéré. Cette proposition montre le caractère intrinsèque du second critère. Par opposition, nous avons mis en évidence (proposition 3) et illustré (proposition 4) une impossibilité à priori pour certaines partitions d'être optimales pour le premier critère. En d'autres termes le premier critère est *biaisé* et on préfèrera donc le second au premier. Mais il n'en reste pas moins que les deux critères sont équivalents dans l'ensemble $\mathcal{R}_r$ des partitions pour lesquelles $|R| = r$.

Cette circonstance nous entraîne à nous placer dans l'ensemble $\mathcal{R}_r$ pour juger de l'adéquation d'une classification aux données. En des termes plus précis, étant donné le couple défini par (ω, P) où ω est l'ordonnance donnée et P une partition, éventuellement définie par un algorithme, la partition P est-elle significativement voisine de ω ? c'est-à-dire : la valeur de $|\omega \cap R \times S|$, où R et S sont relatifs à P, est-elle assez petite compte tenu de la distribution de

$$|\omega \cap R \times R|$$

sur $\mathcal{R}_r$ muni d'une mesure uniforme ?

4 — Distribution asymptotique de $|\omega \cap R \times S|$.

La donnée d'une relation binaire symétrique sur E définit sur F une partition en deux classes R et S où R est l'ensemble des paires pour lesquelles la relation est vérifiée. Nous commencerons par étudier la distribution de $|\omega \cap R \times S|$ sur l'ensemble $\mathcal{B}_r$ des relations symétriques sur E pour lesquelles le cardinal de R est égal à r, $\mathcal{B}_r$ étant muni de la mesure uniforme. On a évidemment $\mathcal{R}_r \subset \mathcal{B}_r$.

Commençons par exprimer $|\omega \cap R \times S|$ d'une manière particulière. A chaque relation associons sa fonction indicatrice, $((\epsilon(p)), p \in F)$, qui représente un point de $\{0, 1\}^F$: $\epsilon(p) = 0$ (resp. 1) si la paire p appartient à R (resp. S). La donnée de ω sera traduite par celle d'une fonction ordinale, $k(p)$, qui à chaque point de F associe son rang pour ω.

Désignant par k_i le rang de la ième paire séparée rencontrée en parcourant F de gauche à droite selon ω, on a

$$|\omega \cap R \times S| = \sum_{i=1}^{s} (k_i - i) = \sum_{i=1}^{s} k_i - \frac{s(s+1)}{2} \quad (1)$$

où $s = |S|$.

On peut encore écrire (1) sous la forme

$$\sum_{p \in F} \epsilon(p)\, k(p) - \frac{s(s+1)}{2} \quad (2)$$

A la classe des entiers $\{k(p)\}\, p \in F$, attachons les caractéristiques suivantes :

a) La moyenne $\bar{k} = \frac{1}{f} \frac{f(f+1)}{2} = \frac{f+1}{2}$ où

$$f = |F| \quad (3)$$

b) La variance $m_2(k) = \frac{1}{f} \sum_{i=1}^{f} i^2 - \left(\frac{f+1}{2}\right)^2$

$$= \frac{(f+1)(2f+1)}{6} - \frac{(f+1)^2}{4}$$

$$= (f+1)(f-1)/12.$$

$$\simeq f^2/12\,. \quad (4)$$

c) Le coefficient de Wald et Wolfowitz

$$W(h) = m_h(k)/(m_2(k))^{h/2}\,,$$

où $m_h(k)$ désigne le moment centré d'ordre h de la classe des entiers $\{k(p)\}\, p \in F$ qui est celle des f premiers entiers naturels $1, 2, \ldots, f$.

$$W(h) = \frac{1}{f} \sum_{i=1}^{f} \left(i - \frac{f+1}{2}\right)^h \Bigg/ \left[\frac{1}{f} \sum_{i=1}^{f} \left(i - \frac{f+1}{2}\right)^2\right]^{h/2}$$

Nous allons vérifier que la condition de Wald et Wolfowitz qui exprime que pour f tendant vers l'infini, $W(h)$ reste *fini* pour tout h, entier *fini* supérieur à 3, $(h \geqslant 3)$, est ici satisfaite. En effet le dénominateur de $W(h)$ peut se mettre sous la forme

$$\left[\frac{(f+1)(f-1)}{12}\right]^{h/2} \nearrow \left(\frac{f}{2\sqrt{3}}\right)^h$$

Il en résulte que

$$W(h) \nearrow (\sqrt{3})^h \frac{1}{f} \sum_{i=1}^{f} \left(\frac{2i}{f} - \frac{f+1}{f}\right)^h$$

or

$$-1 < \frac{2i}{f} - \frac{f+1}{f} < 1$$

D'où en valeur absolue

$$|W(h)| < (\sqrt{3})^h \qquad (5)$$

La distribution des valeurs de $\sum_{p \in F} \epsilon(p)\, k(p)$ est celle de sommes portant sur un échantillon d'effectif s tiré sans remise selon une loi uniforme de l'ensemble fini d'entiers $\{1, 2, \ldots, f\}$. Conformément au théorème établi par Wald et Wolfowitz (1944), (Statistical tests based on permutations of the observations, *Ann. Math. Statist.*, 15, 358-372), en tenant compte de (5) on peut énoncer la

4.1 — *Proposition.* $\mathcal{B}_r$ étant muni d'une mesure uniforme, la distribution de $\sum_{p \in F} \epsilon(p)\, k(p)$ est pour f grand, approximativement normale de moyenne $s\bar{k}$ et de variance $s\left(1 - \frac{s}{f}\right) m_2(k)$.

Il en résulte que la distribution de $|\omega \cap R \times S|$ sur $\mathcal{B}_r$ est approximativement normale de moyenne

$$s\bar{k} - s(s+1)/2 = r \,.\, s/2 \quad \text{où} \quad r = |R|$$

et de variance

$$\left[\frac{f}{(f-1)} \times (f+1)(f-1)/12\right] \frac{rs}{f} = rs(f+1)/12 \,.$$

4.2 — *Valeur de l'approximation normale*

On peut calculer les coefficients B_1 et B_2 de Pearson de la distribution de $\sum_{p \in F} \epsilon(p)\, k(p)$ en fonction de ceux β_1 et β_2 de la distribution des entiers $k(p)$. Rappelons que le calcul des moments de la statistique en question est celui d'une moyenne, pour un tirage sans remise, dans une population finie. On obtient

$$B_1 \simeq \beta_1 \times \frac{(1-2\alpha)^2}{\alpha(1-\alpha)} \times \frac{1}{f^2} \quad \text{où} \quad \alpha = \frac{s}{f}$$

$$B_2 \simeq 3 + \beta_2 \times \left(\frac{1}{\alpha(1-\alpha)} - 6\right) \times \frac{1}{f}$$

Exemple : pour $\alpha = 2/3$ et $f = 300$ (ce qui correspond à $n = 25$), on a

$B_1 \simeq \beta_1/18 \times 10^4$, donc très voisin de 0 .

$B_2 \simeq 3 - \beta_2/200$, donc très voisin de 3 .

4.3 — *Distribution asymptotique de* $|\omega \cap R \times S|$ *sur* $\mathcal{R}_r$.

Retenons l'échantillon des valeurs de $\sum_{p \in F} \epsilon(p) k(p)$ pour lesquelles $(\epsilon(p), p \in F)$ correspond à un élément dans $\mathcal{R}_r$. Nous avons déjà montré [cf. lemme b.2 § 3.1] que si P est un élément pris dans $\mathcal{R}_r$ muni d'une mesure uniforme, on a pour tout p de F

$$\Pr\{\epsilon(p) = 1\} = \sum_{i<j} n_i n_j \Big/ \frac{n(n-1)}{2} = \frac{s}{f} \,,$$

et par conséquent

$$\mathrm{Var}(\epsilon(p)) = \frac{s}{f} - \left(\frac{s}{f}\right)^2 = \frac{r \,.\, s}{f}$$

Notre échantillon est engendré de manière systématique indépendamment de ω ; on peut considérer que l'hypothèse de normalité reste valable dans $\mathcal{R}_r$ pour ce qui concerne la distribution de la statistique $\Sigma\, \epsilon(p)\, k(p)$, (on rapprochera une telle technique d'échantillonnage de certaines, automatiques, permettant d'engendrer des nombres au hasard). D'ailleurs l'échantillon respecte rigoureusement la valeur moyenne dans $\mathcal{B}_r$ de la statistique étudiée, [cf. Th. 3.2 et Prop. 1 § 3.3 (1)].

---

(1) Remarquons qu'on peut ici retrouver simplement ce théorème grâce à l'expression de $|\omega \cap R \times S|$ définie ci-dessus. En effet considérons

$$\frac{1}{|\Lambda|} \sum_{P \in \Lambda} \sum_{p \in F} \epsilon(p)\, k(p) \qquad \text{où} \qquad (\epsilon(p), p \in F)$$

est associé à la relation d'équivalence définie par la partition P et où Λ est l'ensemble des partitions d'un type donné. Inversant les signes sommes et appliquant le lemme b.2 (§ 3.1), on obtient

$$\sum_{p \in F} k(p) \frac{1}{|\Lambda|} \sum_{P \in \Lambda} \epsilon(p) = \frac{s}{f} \sum_{p \in F} k(p) = s\bar{k} = s(f+1)/2 \,.$$

4.4 — *Distribution asymptotique de* $|\omega \cap R \times S|$ *sur* Ω

Ω désigne l'ensemble de tous les ordres totaux pouvant être définis sur F (ordonnances sur E), le couple (R, S) est associé à une partition P fixée ici une fois pour toutes. Par une argumentation identique à celle qui a conduit à la proposition 5.1 ci-dessus, on a

Proposition 4.4. Ω étant muni d'une mesure uniforme et (R, S) correspondant à une partition P fixée, la distribution de

$$\left(|\omega \cap R \times S| - \frac{r \cdot s}{2}\right) \Big/ \sqrt{rs(f+1)/12}$$

est pour f grand, normale centrée réduite. Il en résulte

Corollaire 4.4. La distribution de

$$\left(|\omega \cap R \times S| - \frac{r \cdot s}{2}\right) \Big/ \sqrt{rs(f+1)/12}$$

sur $\Omega \times \mathcal{R}_r$ est, pour f grand, normale centrée réduite.

Terminons ce paragraphe en proposant un test de signification. Si ω est une ordonnance donnée sur E et P une partition déterminée par un algorithme, le calcul de

$$(2\,|\omega \cap R \times S| - r \cdot s)/\sqrt{r \cdot s(f+1)/3} \quad ,$$

qui est celui d'une variable aléatoire normale centrée réduite *dans l'hypothèse d'indépendance de* ω *et de* P, permettra de juger si P est significativement voisine de ω. Le seuil de signification devant être déterminé après une longue expérience.

Notons encore qu'une telle statistique permet de détecter les niveaux les plus significatifs d'une hiérarchie de classifications.

5 — Autres critères

Chacun des deux critères précédents s'interprète aisément comme une mesure dans $F \times F$ du complémentaire de la différence symétrique entre ω et ϖ (où ϖ désigne le graphe dans $F \times F$ du préordre sur F représentant la partition P). On a

$$F \times F - \omega \,\Delta\, \varpi = F \times F - (\omega \cap S \times R + \omega' \cap R \times S + \omega' \cap R \times R + \omega' \cap S \times S),$$

où les signes + et — sont ensemblistes et où ω' est le complémentaire de ω. *Si* ω *est un ordre total sur* F, le critère de Benzécri correspond à adopter sur $F \times F$ une mesure μ_P qui dépend de la partition P et qui est définie comme suit :

$$\mu_P(p, q) = 0 \quad \text{si} \quad (p, q) \in R \times R + S \times S$$

$$= 1 \quad \text{sinon}$$

L'égalité $|\omega \cap R \times S| = \frac{1}{2}\,\mu_P(F \times F - \omega \,\Delta\, \varpi)$ provient alors de l'équivalence :

$$(p, q) \in \omega \cap R \times S \iff (q, p) \in \omega' \cap S \times R \ .$$

Quant au critère de W.F. de la Véga, il correspond à adopter sur $F \times F$ la mesure uniforme qui à chaque sommet (p, q) affecte la masse 1 ; en effet on a l'identité

$$|F \times F - \omega \,\Delta\, R \times S| = |F \times F - \omega \,\Delta\, \varpi|$$

qui résulte de l'équivalence

$$(p, q) \in \omega \cap (R \times R + S \times S) \iff (q, p) \in \omega' \cap (R \times R + S \times S) \ ,$$

ω étant un ordre total.

Relativement au point de vue présenté au paragraphe précédent, la quantité critère de la Véga s'obtient à partir de celle, $|\omega \cap R \times S|$, de Benzécri en centrant $|\omega \cap R \times S|$. L'étude précédente nous incite à réduire la variable centrée et à adopter comme critère de proximité

$$\frac{|\omega \cap R \times S| - \dfrac{r \cdot s}{2}}{\sqrt{r \cdot s}} \qquad (1)$$

Exemple : Reprenons les partitions (I) et (II) de l'exemple précédent (§ 2). Les valeurs de (1) sont respectivement pour (I) et (II) :

$$25/\sqrt{6 \times 9} = 25/\sqrt{54} = 3{,}40$$

et $20/\sqrt{7 \times 8} = 20/\sqrt{56} = 2{,}67$. La préférence de la partition (I) se trouve avec ce critère accentuée par rapport au critère $|\omega \cap R \times S| - r \,.\, s/2$.

Toutefois, comme le souligne d'ailleurs J.P. Benzécri : "quelle que soit notre confiance dans un critère, il reste que la recherche d'une classification doit tenir compte de tant d'exigences contradictoires, qu'il serait présomptueux de prétendre avoir, d'une seule formule, défini l'optimum absolu".

Nous allons dans ces conditions proposer différents critères qui, à défaut de définir des préordres totaux sur l'ensemble des partitions, définissent des ordres (ou préordres) partiels.

5.1 — *Critère de l'ordre lexicographique*

a) *Définitions préliminaires*

1/ Relativement à un préordre total ω sur un ensemble fini F, nous définirons le rang d'un élément p de F, $\rho(p)$, comme étant le cardinal

$$\rho(p) = |\{q \in F : q \leqslant p\}| \quad .$$

2/ Le type d'un préordre total sur F sera la suite des cardinaux de ses différentes classes de gauche à droite.

3/ ω étant un préordre total *fixé* sur F, ϖ et ϖ' deux préordres totaux *de même type* sur F ; on posera

$$\varpi < \varpi' \text{ (lexico. } \omega)$$

s'il existe une section C strictement commençante pour ω telle que :

(i) $\rho(p) \leqslant \rho'(p)$ pour tout p de C, ρ et ρ' désignant respectivement les rangs de p pour ϖ et ϖ' ;

(ii) il existe un p de C pour lequel on ait strictement $\rho(p) < \rho'(p)$.

b) *Critère proprement dit*

Reprenant nos notations où ω désigne le préordre donné sur l'ensemble F des paires d'objets distincts de E ; si ω et ω' sont les préordres sur F associés à deux partitions d'un même $\mathcal{R}_r = \{P : |R| = r\}$, ϖ et ϖ' sont de même type (r, s) et on préfèrera ϖ' à ϖ si

$$\varpi < \varpi' \text{ (lexico. } \omega) \qquad (2)$$

Exemple : considérons sur l'ensemble

$$E = \{a, b, c, d, e, f\}$$

la préordonnance de base $\overline{\omega}$ définie dans l'exemple du paragraphe 2 ci-dessus ; soit

$$ad = bc < be = af = df < de < bd = ce = ef < ab = \\ = ac = cd < ae = bf < cf$$

où on a noté xy la paire $\{x, y\}$.

Soit à comparer vis-à-vis du critère que l'on vient de définir, les deux partitions

$$\{\{a, d, f\}, \{b, c, e\}\} \qquad (I)$$

et

$$\{\{a, d, e, f\}, \{b\}, \{c\}\} \qquad (II)$$

pour lesquelles $|R|$ est le même, soit 6.

Le préordre ϖ associé à (I) est le suivant

$$ad = bc = be = af = df = ce < de = bd = ef = ab = \\ = ac = cd = ae = bf = cf \quad ,$$

celui ϖ' associé à (II) étant

$$ad = af = df = de = ef = ae < bc = be = bd = ce = \\ = ab = ac = cd = bf = cf \quad .$$

Considérons la section $C = [ad, bc]$ strictement commençante de $\overline{\omega}$, on a avec les notations ci-dessus :

$$\rho(ad) = \rho'(ad)$$
$$\rho(bc) < \rho'(bc)$$

donc la première partition est préférable à la seconde.

5.2 — *Critère du nombre d'inversions*

a) Définissons relativement au graphe ω

$$\omega_1 = \{(p, q) : ((p, q) \in \omega) \text{ et } \neg((q, p) \in \omega)\}$$

et

$$\omega_2 = \{(p, q) : (q, p) \in \omega_1\}$$

b) Dans un même ensemble de partitions $\mathcal{R}_r$, une partition sera d'autant plus préférable que

$$|\omega_1 \cap S \times R + \omega_2 \cap R \times S| = 2|\omega_1 \cap S \times R| = \\ = 2|\omega_2 \cap R \times S|, \quad (3)$$

est plus petit.

Le critère d'ordre partiel défini ici est, dans sa restriction à chacun des ensembles $\mathcal{R}_r$, un préordre total.

Exemple : reprenons l'exemple ci-dessus ; pour la partition (I) ce critère vaut 2 et pour la partition (II) $2(1 + 1 + 2 + 2 + 7) = 26$. Nous avons effectué le décompte de $|\omega_2 \cap R \times S|$ en déterminant pour chaque élément p de R le nombre d'éléments q se trouvant strictement à gauche de p pour $\overline{\omega}$ et appartenant à S.

Remarquons que si ω est un ordre total, la quantité critère (3) se met sous la forme $2|\omega \cap S \times R|$ qu'on aura tendance à vouloir comparer avec la quantité $|\omega \cap R \times S|$. Mais, même dans un ensemble $\mathcal{R}_r$, maximiser $|\omega \cap R \times S|$ ne revient pas exactement à minimiser $|\omega \cap S \times R|$. Sous le point de vue exprimé au paragraphe précédent, la distribution de $|\omega \cap S \times R|$ a les mêmes caractéristiques que celle de $|\omega \cap R \times S|$ et on aurait, bien sûr, pu définir un critère tel que (1) en centrant et réduisant la variable $|\omega \cap S \times R|$. Mais, ici, nous préférons garder au critère son caractère partiel et insister sur le fait qu'en général ω est un *préordre* et non un ordre total. Toutefois nous comparerons ce critère à celui d'ordre lexicographique dans le cas où ω est un ordre total et dans le cadre plus général de la recherche d'une chaîne de partitions.

5.3 — *Nouveau critère de J.P. Benzécri — W.F. de la Véga*

Benzécri reste plus exigeant que nous-mêmes, il demande à son critère d'être sûr en se préoccupant peu de son caractère partiel : "... on cherchera un ordre partiel, qui ne donne pas toujours entre deux classifications une préférence (ou une équivalence) mais qui n'exprime que des préférences avec lesquelles on soit comme forcé de s'accorder, et qui en exprime le plus grand nombre possible. Cet ordre partiel nous est fourni par un principe de dominance énoncé par W.F. de la Véga (1965)..." [cf. [3, (b)]].

La formulation la plus simple de ce critère appelé d'ordre latéral peut être la suivante :

P et P' étant deux partitions d'un même ensemble $\mathcal{R}_r$, R et S (resp. R' et S') les classes du préordre associé à P (resp. P') ; on préfèrera P' à P s'il existe une bijection φ de R sur R' telle

$$(\forall p \in R) \quad \varphi(p) \leqslant p \quad \text{pour } \omega .$$

On notera dans ce cas Dr. (P', P)

Exemple : toujours avec le même exemple que ci-dessus, considérons l'application φ de R_2 sur R_1 (où R_1 est associé à la partition (I) et R_2 à (II)) définie comme suit :

p	ad	af	df	de	ef	ae
$\varphi(p)$	ad	af	df	bc	be	ce

on a bien ici $\varphi(p) \leqslant p$ pour $\overline{\omega}$; au vu de $\overline{\omega}$ la partition (I) est meilleure que (II).

Il est rassurant de se rendre compte que, *dans un même ensemble $\mathcal{R}_r$ de partitions, si P' est préféré à P du point de vue de l'ordre latéral, il l'est à fortiori du point de vue de l'un quelconque des critères précédents.* Un point optimal dans $\mathcal{R}_r$ pour l'un de ces critères est donc un point extrémal pour la restriction de Dr. à $\mathcal{R}_r$.

Il existe un algorithme très simple permettant de définir dans certains ensembles $\mathcal{R}_r$ des partitions optimales pour le critère de l'ordre lexicographique. Ces partitions sont des points extrémaux pour l'ordre latéral ; mais, en général, on voit mal comment explorer l'ensemble de tous les points extrémaux dans un même $\mathcal{R}_r$ pour Dr. ; et pour peu que l'aptitude de l'ensemble E à être classifié *dans* $\mathcal{R}_r$ ne soit pas très grande, on peut se retrouver devant un ensemble de points extrémaux trop vaste pour s'y retrouver.

Il reste, comme le comprenait d'ailleurs de la Véga, que ce critère représente une sorte de "garde fou" pour tout critère défini ou à définir.

B — RECHERCHE D'UNE CHAINE DE PARTITIONS

1 — Représentation d'une chaîne de partitions

La donnée d'une chaîne de partitions sera caractérisée par celle d'un préordre total sur F, satisfaisant certaines conditions, qu'on appellera "préordonnance ultramétrique".

1.1 — *Définition*

Une préordonnance ϖ sur E est dite ultramétrique si la condition suivante est satisfaite :

$$(\forall x, y \text{ et } z \in E) \quad \rho(x, y) \leqslant r \text{ et } \rho(y, z) \leqslant r \Longrightarrow \rho(x, z) \leqslant r$$

où r est un entier donné et $\rho(x, y)$ désigne le rang de la paire $\{x, y\}$ pour ϖ.

Les termes de la définition ci-dessus se justifient par la

1.2 — *Proposition*

La condition nécessaire et suffisante pour qu'une distance d, définie sur E, soit ultramétrique est que la préordonnance associée le soit.

Rappelons que la préordonnance sur E, associée à une distance d, est définie de la manière suivante :

$$\{x, y\} \leqslant \{z, t\} \Longleftrightarrow d(x, y) \leqslant d(z, t)$$

La condition est nécessaire. En effet si d est ultramétrique

$$d(x, z) \leqslant \max(d(x, y), d(y, z)) \quad \text{pour tout } x, y \text{ et } z \text{ de } E, \quad (1)$$

une telle condition implique, relativement au préordre associé,

$$\rho(x, z) \leqslant \max(\rho(x, y), \rho(y, z)) \quad \text{pour tout } x, y \text{ et } z \text{ de } E, \quad (2)$$

qui entraîne manifestement la condition de l'énoncé.

Réciproquement, si la préordonnance associée à d est ultramétrique on a la condition (2) qui implique (1).

1.3 — *Représentation*

Les différentes sections strictement commençantes d'une préordonnance ultramétrique ϖ sont saturées ; nous voulons dire que si R est une telle section

$$\{x, y\} \in R \text{ et } \{y, z\} \in R \Longrightarrow \{x, z\} \in R$$

Or la donnée d'une chaîne de partitions est équivalente à la donnée d'une suite croissante (au sens de l'inclusion) de parties de F saturées. Si

$$R_1 \subset R_2 \subset \ldots \subset R_k \subset \dot{R}_{k+1} = F$$

définit une telle suite, R_i est l'ensemble des paires réunies par la ième partition de la chaîne : P_i. P_i est plus fine que P_{i+1}, et avec les notations du chapitre 0, ($P_i < P_{i+1}$). De la sorte les différentes sections commençantes du préordre ultramétrique associé à la chaine de partition sont respectivement

$$R_1, R_2, \ldots, R_k \text{ et } F .$$

On peut aussi définir ce préordre au moyen d'une fonction ordinale qui à chaque paire p de F associe le plus petit indice i tel que $p \in R_i$. Le nombre de partitions d'une même chaîne étant au plus n, on a $k \leqslant n - 2$.

L'ensemble représenté et celui représentant sont ici équipotents (dans notre cas de même cardinal). C'est là une qualité de notre représentation ; une préordonnance ultramétrique est en fait la classe d'équivalence formée par l'ensemble des distances ultramétriques qui définissent une même chaîne de partitions.

2 — Généralisation des critères précédents

2.1 — *Critère initial de Benzécri*

Comme nous l'avons fait pour un préordre quelconque ω [cf. § A 5.2.a] associons à une préordonnance ultramétrique ϖ

$\varpi_1 = \{(p\,,q) : ((p\,,q) \in \varpi)$ et $\neg((q\,,p) \in \varpi)\}$

et

$$\varpi_2 = \{(p\,,q) : (q\,,p) \in \varpi_1\}$$

ω étant l'ordonnance donnée, on généralisera, pour la recherche d'une chaîne de partitions, le critère de J.P. Benzécri par $|\omega \cap \varpi_1|$. Si

$$(R_1\,, R_2\,, \ldots\,, R_k\,, F)$$

est la suite des sections commençantes de F relativement à ω, en posant :

$$T_1 = R_1\,, \ldots\,, T_i = R_i - R_{(i-1)}\,, \ldots\,, T_{k+1} = F - R_k\,,$$

on a :

$$|\omega \cap \varpi_1| = \left| \omega \cap \sum_{i=1}^{k} T_i \times \sum_{j>i} T_j \right|$$

2.2 — *Critère de la différence symétrique*

Au sens de ce critère, on préférera d'autant plus une chaîne de partitions que $|\omega \Delta \varpi|$ est plus petit ; ω désignant toujours le préordre donné sur F et ϖ la préordonnance ultramétrique associée à la chaîne de partitions.

Nous allons à présent généraliser les différents critères d'ordre partiel définis ci-dessus.

2.3 — *Critère d'ordre lexicographique*

Reprenant les définitions préliminaires du paragraphe A.5.1, si ϖ et ϖ' sont deux préordonnances ultramétriques de *même type*, on préférera la chaîne de partitions définie par ϖ' à celle définie par ϖ si

$$\varpi < \varpi' \quad \text{(lexico. } \omega\text{)}$$

où ω est le préordre donné sur F.

Dire que ϖ et ϖ' sont de même type revient ici à dire que les différentes sections commençantes, $R_1\,, R_2\,, \ldots\,, R_k\,, F$, définies par ϖ ; et celles,

$$R'_1\,, R'_2\,, \ldots\,, R'_k\,, F\,,$$

par ϖ' sont respectivement de même cardinal

$$|R_i| = |R'_i|$$

pour tout i.

2.4 — *Critère du nombre d'inversions*

Dans l'ensemble Γ des préordres ultramétriques d'un *type donné*, une chaîne de partitions, ϖ, est d'autant plus préférable que $|\omega_1 \cap \varpi_2 + \omega_2 \cap \varpi_1|$ est plus petit. ω_1 et ω_2 sont associés à ω de la même façon que ϖ_1 et ϖ_2 à ϖ.

2.5 — *Critère d'ordre latéral*

ϖ et ϖ' représentant deux chaînes de partitions dans l'ensemble des préordres ultramétriques d'un type donné, on préférera ϖ' à ϖ s'il existe une bijection φ de F sur lui-même telle que

(i) la restriction de φ à toute section strictement commençante de ϖ est une section strictement commençante de ϖ' ; c'est-à-dire, avec des notations familières : $\varphi(R_i) = R'_i$ pour tout $i = 1\,, 2\,, \ldots\,, k$;

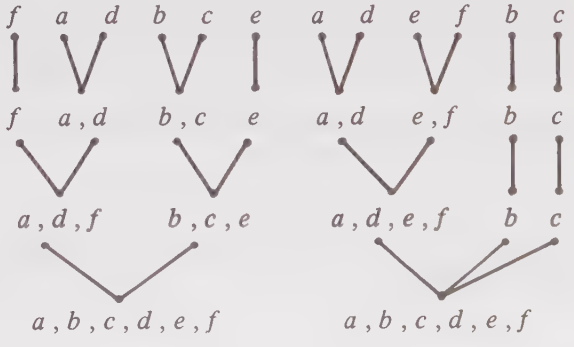
(ii) $(\forall_i = 1\,, 2\,, \ldots\,, k)$ et $(\forall p \in R_i)$ $\varphi(p) \leqslant p$ pour ω. On notera Dr $(\varpi'\,, \varpi)$.

Ici encore, relativement au critère d'ordre latéral, nous pouvons tirer les mêmes conclusions que celles formulées au paragraphe 5.3 précédent.

2′) *Exemples* : au vu de la préordonnance $\underline{\omega}$

$$af < ad = df < bc < ac = ef < ab = bd = cd = be = \\ = ce = de = bf < cf = ae\,,$$

où on a noté xy la paire $\{x\,, y\}$, établie à partir du tableau d'incidence des données (Ch. 1, § II.2), sur la base de l'indice s, nous allons comparer, du point de vue de chacun des cinq critères qu'on vient de définir, les deux chaînes de partitions :



(I) (II)

La préordonnance ultramétrique associée à (I), ϖ_1, est :

$$ad = bc < af = df = be = ce < ab = ac = ae = bd = bf = cd = cf = de = ef \quad .$$

et celle associée à (II), ϖ_2 :

$$ad = ef < ae = af = de = df < ab = ac = bc = bd = be = bf = cd = ce = cf \quad .$$

ϖ_1 et ϖ_2 sont de même type (2, 4, 9).

Soit le tableau des rangs

	af	*ad*	*df*	*bc*	*ac*	*ef*	*ab*	*bd*	*cd*	*be*	*ce*	*de*	*bf*	*cf*	*ae*
$\underline{\omega}$	1	3	3	4	6	6	13	13	13	13	13	13	13	15	15
ϖ_1	6	2	6	2	15	15	15	15	15	6	6	15	15	15	15
ϖ_2	6	2	6	15	15	2	15	15	15	15	15	6	15	15	6

rappelons que le rang d'un élément p de F, $\rho(p)$, est le cardinal

$$\rho(p) = |\{q \in F : q \leqslant p\}|$$

2'1) *Critère initial de Benzécri*

Avec les notations précédentes ce critère se met sous la forme

$$\sum_{i=1}^{k} \left| \omega \cap T_i \times \sum_{j>i} T_j \right|$$

on obtient pour ϖ_1 :

$$12 + 11 + 9 + 9 + 7 + 7 = 55 \ ;$$

pour chaque élément p de T_i nous avons calculé le cardinal de l'intersection de $\sum_{j>i} T_j$ et de la section finissante (au sens large) pour $\underline{\omega}$ ($q \geqslant p$, $q \neq p$).

Pour la préordonnance ultramétrique ϖ_2, on obtient : $12 + 10 + 1 + 9 + 7 + 9 = 48$.

La hiérarchie de classifications (I) est donc préférable à (II).

2'2) *Critère de la différence symétrique*

Nous déterminons la valeur de ce critère pour ϖ_i à partir du tableau des rangs ci-dessus de la façon suivante : considérons les éléments de F de gauche à droite pour *l'ordre où ils se présentent dans le tableau* ; si p est un élément courant appartenant à F, couplons p avec chacun des éléments de F à sa droite, soit q un tel élément. La contribution de la paire $\{p , q\}$ au cardinal de la différence symétrique $|\underline{\omega} \Delta \varpi_i|$, peut être 0, 1 ou 2. 0 si p et q se présentent exactement dans le même ordre pour $\underline{\omega}$ et ϖ_i, 1 si pour l'un des deux préordres on a $p < q$ ou $p > q$ et pour l'autre $p = q$, enfin 2 s'il y a une inversion stricte, c'est-à-dire si pour l'un des deux préordres $p < q$ et pour l'autre $p > q$.

Pour ϖ_1, la valeur du critère est :

$$7 + 1 + 4 + 0 + 11 + 11 + 2 + 2 + 2 + 4 + 4 = 48 \ .$$

Pour ϖ_2, la valeur du critère est :

$$7 + 2 + 4 + 14 + 12 + 0 + 4 + 4 + 4 + 4 + 4 + 2 + 3 + 1 = 65 \ ;$$

même conclusion que ci-dessus.

2'3) *Critère d'ordre lexicographique*

Considérons la section $C = [af , ad , df , bc]$ strictement commençante pour $\underline{\omega}$, on a avec les notations du paragraphe A.5.1.

$$\rho_1(af) = \rho_2(af) = 6$$
$$\rho_1(ad) = \rho_2(ad) = 2$$
$$\rho_1(df) = \rho_2(df) = 6$$

et

$$\rho_1(bc) = 2 < \rho_2(bc) = 15$$

même conclusion que précédemment.

2'4) *Critère du nombre d'inversions*

Le calcul peut être effectué à partir du tableau des rangs d'une manière analogue à 2).

Pour ϖ_1 la valeur du critère est :

$$2 + 0 + 1 + 0 + 2 + 2 + 0 = 7$$

Pour ϖ_2 la valeur du critère est :

$$2 + 0 + 1 + 3 + 2 + 0 + 1 + 1 + 1 + 1 + 1 + 1 = 14$$

La chaîne de partitions (I) est ici encore préférable à (II).

2'5) *Critère d'ordre latéral*

Avec les notations de 2.5, $k = 2$; il y a pour ϖ_1 (resp. ϖ_2) deux sections strictement commençantes. Pour ϖ_1

$$R_1 = [ad\,,\,bc] \quad , \quad R_2 = [af\,,\,df\,,\,be\,,\,ce]\,;$$

Pour ϖ_2

$$R'_1 = [ad\,,\,ef] \quad , \quad R'_2 = [ae\,,\,af\,,\,de\,,\,df]\,.$$

Considérons l'application φ définie comme suit :

p	ad	bc	af	df	be	ce
$\varphi(p)$	ad	ef	af	df	de	ae

on a :

$$R'_1 = \varphi(R_1) \quad \text{et} \quad R'_2 = \varphi(R_2)\,;$$

de plus

$$(\forall p \in R_1 \cup R_2)\ \ \varphi(p) \leqslant p \quad \text{pour} \quad \underline{\omega}\,.$$

Donc le critère d'ordre latéral avantage la chaîne (I) par rapport à la chaîne (II). Il n'est donc pas étonnant que (I) ait été préféré à (II) du point de vue de l'un quelconque des critères précédents.

3 — Rang lexicographique et nombre d'inversions

Comme nous l'avions annoncé, nous allons comparer les deux critères, d'ordre lexicographique et du nombre d'inversions dans le cas où la donnée ω *est un ordre total sur F.*

3.1 — *Position du problème*

Pour répondre à la question suivante que nous formulons de manière intuitive : "Si ϖ est un préordre ultramétrique sur F, assez proche de ω, au sens du critère (2.3), en est-il de même au sens du critère (2.4) ?", il s'agira pour nous de comparer la proximité au sens du nombre d'inversions à celle au sens du critère d'ordre lexicographique. Nous serons amenés à le faire, (parce que ω est un ordre total), sur l'ensemble Ω des ordres totaux sur F. En effet, si ϖ est un préordre total, (en l'occurrence ultramétrique), sur F, faisons-lui correspondre, l'ordre total $o(\varpi)$, compatible avec ϖ, le plus voisin de ω, obtenu en adoptant sur chaque classe de ϖ, la restriction de ω à la partie de F que constitue la classe. Il est important de noter, dans ce cadre, que si ϖ et ϖ' sont deux préordonnances ultramétriques de même type, on a $\varpi < \varpi'$ (lexico. ω) $\Longleftrightarrow o(\varpi) < o(\varpi')$ (lexico. ω). D'autre part le nombre d'inversions que présente ϖ par rapport à ω est exactement celui que présente $o(\varpi)$ par rapport à ω. Compte tenu de la généralité de la question, nous nous abstrairons dans la suite de ce que représente F.

3.2 — *Rang lexicographique*

Soit donc un ensemble fini F de cardinal f et ω un ordre total fixé sur F. Nous désignerons les éléments de F par leurs rangs pour ω. La donnée d'un ordre total o sur F est équivalente à celle d'une permutation $(o(1)\,,\,o(2)\,,\,\ldots\,,\,o(f))$ où $o(i)$ est l'objet de F désigné par le rang de l'élément i pour o ; de la sorte, nous parlerons indifféremment d'ordre total ou de permutation sur F. Ici, la relation d'ordre lexicographique devient : $o < o'$ (lexico ω) si et seulement s'il existe i de F tel que

$$(\forall j < i)\ ;\ o(j) = o'(j) \qquad (1)$$

$$o(i) < o'(i) \quad \text{pour} \quad \omega\,. \qquad (2)$$

Cette relation d'ordre lexicographique est totale sur l'ensemble des permutations, Ω, de F. Nous avons, dans ce cadre, établi un algorithme qui permet de

définir le successeur lexicographique immédiat d'une permutation donnée [cf. Annexe suivante].

3.2.1 — *Définition*

Le rang lexicographique d'un ordre o est le cardinal $|]\omega , o]|$, où l'intervalle $]\omega , o]$ est défini sur Ω muni de la relation d'ordre lexicographique.

3.3 — *Nombre d'inversions*

Le nombre d'inversions entre deux ordres totaux sur F est le cardinal de la différence symétrique de leurs graphes respectifs dans $F \times F$.

Nous noterons $\rho(o)$ le rang lexicographique de l'ordre total o et $\nu(o)$ le nombre d'inversions que présente o par rapport à ω. L'objet du présent paragraphe peut ici être précisé : l'application $o \longrightarrow \rho(o)$ étant bi-univoque, si $\rho(o) = \rho$ donné, déterminer $\nu(o)$; en d'autres termes définir l'application $\rho \longrightarrow \nu$.

3.3.1 — *Lemme*

L'application $\rho \longrightarrow \nu(\rho)$ est bien définie.

Soient :

$(o(1), o(2), \ldots, o(f))$ et $(o'(1), o'(2), \ldots, o'(g))$ deux permutations respectives de $(1, 2, \ldots, f)$ et $(1, 2, \ldots, g)$ qui ont même rang lexicographique ; le nombre d'inversions que présente

$$(o(1), o(2), \ldots, o(f))$$

par rapport à $(1, 2, \ldots, f)$ est le même que celui que présente $(o'(1), o'(2), \ldots, o'(g))$ par rapport à $(1, 2, \ldots, g)$, quels que soient les entiers f et g. On ne compromet en rien la généralité si on suppose $g > f$; soit $g = f + k$. On a alors nécessairement :

$$(o'(1), o'(2), \ldots, o'(k), o'(k+1), \ldots, o'(g)) =$$
$$= (1, 2, \ldots, k, o(1) + k, o(2) + k, \ldots, o(f) + k)$$

c'est-à-dire : $o'(l) = l$ pour $l \leqslant k$

$$o'(k + m) = k + o(m) \text{ pour } l \leqslant m \leqslant f.$$

Il est manifeste que la permutation

$$(o(1), o(2), \ldots, o(f))$$

présente par rapport à $(1, 2, \ldots, f)$ le même nombre d'inversions que

$$(1, 2, \ldots, k, o(1) + k, \ldots, o(f) + k)$$

par rapport à

$$(1, 2, \ldots, k, 1 + k, 2 + k, \ldots, f + k) .$$

3.3.2 — *Théorème*

Si ρ est le rang lexicographique de la permutation $(o(1), o(2), \ldots, o(f))$ et si s est l'entier pour lequel $s! \leqslant \rho < (s+1)!$, considérons la suite des divisions

$$\begin{aligned}
\rho &= q \times s! + \rho_1 , && \rho_1 < s! \\
\rho_1 &= q_1 \times (s-1)! + \rho_2 , && \rho_2 < (s-1)! \\
\rho_2 &= q_2 \times (s-2)! + \rho_3 , && \rho_3 < (s-2)! \\
&\ldots\ldots\ldots\ldots\ldots\ldots \\
\rho_{(s-1)} &= q_{(s-1)} \times 1! + 0 .
\end{aligned}$$

On a :

$$\nu(\rho) = q + q_1 + q_2 + \ldots + q_{(s-1)} .$$

Preuve

On a $\rho < k!$ si est seulement si la section commençante de $o(o(1), o(2), \ldots, o(f-k))$ est

$$(1, 2, \ldots, (f-k)) ;$$

en effet, il y a $k!$ permutations dont la section commençante est $(1, 2, \ldots, f-k)$ dont la première est celle, $(1, 2, 3, \ldots, f)$, de rang 0 et la dernière, $(1, 2, \ldots, f-k, f, f-1, f-2, \ldots, f-k+1)$, de rang $(k! - 1)$.

Il en résulte que $s! \leqslant \rho < (s+1)!$ si et seulement si la section commençante de o, de cardinal $f-s$, se présente sous la forme

$$\sigma = (1, 2, \ldots, f-s-1, o(f-s)) ,$$

où $o(f-s)$ est différent de $(f-s)$, d'où nécessairement : $o(f-s) > (f-s)$. Il y a $s!$ permutations pour lesquelles σ est une section commençante avec $o(f-s) = f-s$; la dernière d'entre elles ayant pour rang $(s! - 1)$. Il y a $2\,s!$ permutations pour lesquelles σ est une section commençante avec $(f-s) \leqslant o(f-s) \leqslant (f-s+1)$; la dernière d'entre elles ayant pour rang $(2s! - 1)$.

Plus généralement, il y a $(k+1) \times s!$ permutations pour lesquelles σ est une section commençante avec

$$(f-s) \leqslant o(f-s) \leqslant (f-s+k)$$

la dernière d'entre elles ayant pour rang

$$(k+1)s! - 1 \quad .$$

D'où : Pour que le quotient de la division de ρ par $s!$ soit q, il est nécessaire et suffisant que $o(f-s) = f-s+q$, o se présente donc nécessairement sous la forme :

$$o = (1, 2, \ldots, f-s-1, f-s+q, o(f-s+1), o(f-s+2), \ldots, o(f))$$

et le reste de la division ρ_1 est précisément le rang lexicographique de

$$\theta = (o(f-s+1), o(f-s+2), \ldots, o(f))$$

par rapport à

$$\dot{\tau} = ((f-s), (f-s+1), \ldots, (f-s+q-1), (f-s+q+1), \ldots, f) .$$

Le nombre d'inversions que présente o par rapport à $\omega = (1, 2, \ldots, f)$ est celui que présente le terme $(f-s+q)$ plus celui que présente θ par rapport à τ. Pour o, l'élément

$$o(f-s) = (f-s+q)$$

présente q inversions avec les termes suivants parmi lesquels se trouvent, en effet,

$$(f-s), (f-s+1), \ldots, (f-s+q-1) .$$

Par ailleurs, le nombre d'inversions que présente o par rapport à τ est exactement $\nu(\rho_1)$.

De proche en proche on aboutit au résultat annoncé.

3.3.3 — *Corollaire*

$$\rho < k! \Longrightarrow \nu(\rho) \leqslant k(k-1)/2 .$$

Ce résultat se déduit immédiatement du premier paragraphe de la démonstration précédente. En effet, dans l'ensemble des permutations pour lesquelles $\rho < k!$, celle pour laquelle $\nu(\rho)$ est le plus élevé est celle de rang $\rho = (k! - 1)$, soit

$$(1, 2, \ldots, f-k, f, f-1, f-2, \ldots, f-k+1) .$$

Reprenons ici l'interprétation de F comme ensemble des paires d'objets distincts de E, de ω comme celle de l'ordonnance donnée et de o comme étant l'ordre total sur F associé à une préordonnance ϖ ultramétrique, $o = o(\overline{\omega})$. Le corollaire ci-dessus répond à la question posée au paragraphe 3.1 précédent et exprime que si le rang lexicographique de $o(\varpi)$ est inférieur à un seuil donné, il en est de même du nombre d'inversions.

3′ — Annexe au paragraphe précédent

Nous allons ici nous écarter du sujet principal pour illustrer le théorème 3.3.2 précédent. Soit ω un ordre total sur un ensemble fini F de cardinal f et

$$[r(1), r(2), \ldots, r(n)] \quad .$$

la permutation correspondante, nous nous proposons de définir à l'aide d'un algorithme la permutation qui suit immédiatement. Nous commencerons par exposer l'algorithme puis montrerons qu'il définit bien le successeur immédiat d'une permutation.

a) *Algorithme*

Soit donc la permutation

$$[r(1), r(2), \ldots, r(f)] \qquad (1)$$

Désignons par i le plus grand indice pour lequel $r(i-1) < r(i)$; de la sorte, la suite des numéros $r(f), r(f-1), \ldots, r(i)$ est strictement croissante. Dans cette suite croissante, indiquons par $(i+h)$ le plus grand indice pour lequel $r(i+h)$ est plus grand que $r(i-1)$, $(r(i+h) > r(i-1))$; de la sorte $r(i+h)$ est, parmi $r(f), r(f-1), \ldots, r(i)$, le plus petit numéro qui soit plus grand que $r(i-1)$.

Transposons $r(i-1)$ et $r(i+h)$, la permutation initiale devient :

$$r(1), r(2), \ldots, r(i-2), r(i+h), r(i), r(i+1), \ldots, r(i+h-1), r(i-1), r(i+h+1), \ldots, r(f) . \qquad (2)$$

On a, de par la définition de $r(i+h)$:

$$r(i+h-1) < r(i-1) < r(i+h+1) .$$

Il en résulte que la suite des numéros

$$(*)\ r(i), r(i+1), \ldots, r(i+h-1), r(i-1), r(i+h+1), \ldots, r(f)$$

est encore strictement décroissante.

Effectuons sur cette section finissante (*) la permutation inverse, (2) devient

$$r(1), r(2), \ldots, r(i-2), r(i+h), r(f), \ldots, r(f-1), \ldots, r(i+h+1). r(i-1), r(i+h-1), \ldots, r(i) \quad . \qquad (3)$$

La section finissante $r(f), \ldots, r(i)$ est de la sorte strictement croissante.

b) *Proposition*

La permutation (3) définie ci-dessus est le successeur lexicographique immédiat de la permutation initiale (1).

Considérons la permutation donnée

$$r(1), r(2), \ldots, r(i-1), r(i), \ldots, r(i-h), \ldots, r(f) \quad . \qquad (1)$$

La section finissante $r(i), r(i+1), \ldots, r(f)$ étant décroissante

$$r(i) > r(i+1) > \ldots > r(f) .$$

La permutation (1) est la plus grande (pour l'ordre lexicographique) parmi celles pour lesquelles la section commençante $(r(1), r(2), \ldots, r(i-1)$ est fixée. Donc tout successeur de (1) diffère de (1) sur l'un au moins des éléments de cette section commençante.

Toute permutation dont la section commençante est $r(1), r(2), \ldots, r(i-2)$ est plus proche de (1) que toute celle pour laquelle ce n'est pas le cas. Donc le successeur immédiat de (1) est tel que la première composante où il diffère de (1) est précisément $(i-1)$. La composante $(i-1)$ du successeur cherché est à prendre dans la section finissante $r(i), r(i+1), \ldots, r(f)$; c'est nécessairement, parmi les numéros plus grands que $r(i-1)$, celui le plus petit ; soit $r(i+h)$.

Le successeur immédiat a donc bien comme section commençante $r(1), r(2), \ldots, r(i-2), r(i+h)$.

Les $(f-i+1)$ autres éléments sont ordonnés d'une manière croissante dans la section finissante de la permutation (3) qui précède donc toute permutation de section commençante

$$r(1), r(2), \ldots, r(i-2), r(i+h) .$$

Nous allons à l'aide de l'algorithme (a) engendrer les $5 ! = 120$ permutations de $(1, 2, 3, 4, 5)$ ordonnées lexicographiquement et représenter sur cet exemple la fonction $\rho \longrightarrow \nu(\rho)$.

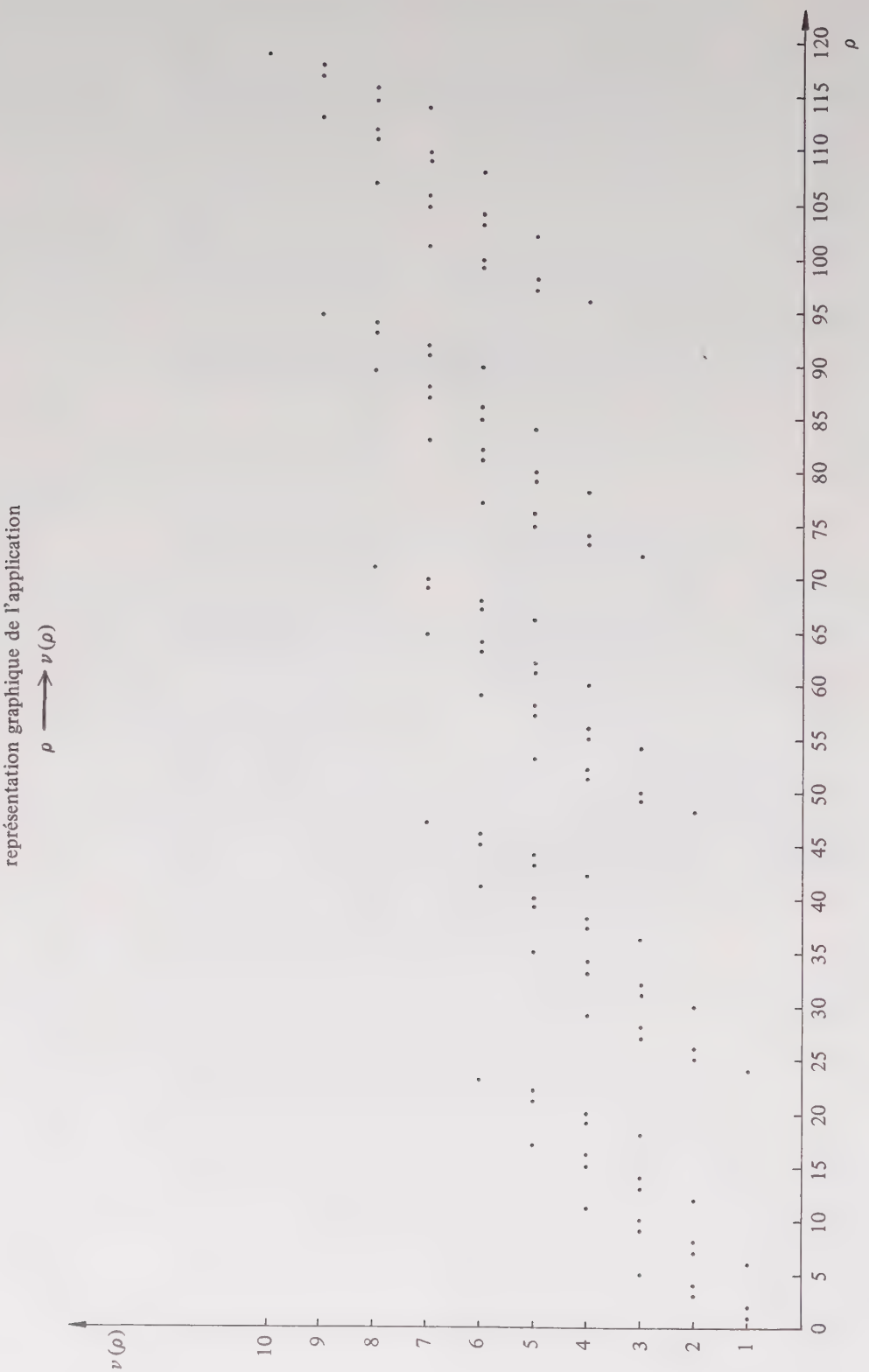
1	2	3	4	5	ρ	$\nu(\rho)$
1	2	3	4	5	0	0
1	2	3	5	4	1	1
1	2	4	3	5	2	1
1	2	4	5	3	3	2
1	2	5	3	4	4	2
1	2	5	4	3	5	3
1	3	2	5	4	6	1
1	3	4	2	5	7	2
1	3	4	5	2	8	2
1	3	5	2	4	9	3
1	3	5	4	2	10	3
1	4	2	3	5	11	4
1	4	2	5	3	12	2
1	4	3	2	5	13	3
1	4	3	5	2	14	3
1	4	5	2	3	15	4
1	4	5	3	2	16	4
1	5	2	3	4	17	5
1	5	2	4	3	18	3
1	5	3	2	4	19	4
1	5	3	4	2	20	4
1	5	4	2	3	21	5
1	5	4	3	2	22	5
1	5	4	3	2	23	6

2	1	3	4	5	ρ	$\nu(\rho)$
2	1	3	4	5	24	1
2	1	3	5	4	25	2
2	1	4	3	5	26	2
2	1	4	5	3	27	3
2	1	5	3	4	28	3
2	1	5	4	3	29	4
2	3	1	4	5	30	2
2	3	1	5	4	31	3
2	3	4	1	5	32	3
2	3	4	5	1	33	4
2	3	5	1	4	34	4
2	3	5	4	1	35	5
2	4	1	3	5	36	3
2	4	1	5	3	37	4
2	4	3	1	5	38	4
2	4	3	5	1	39	5
2	4	5	1	3	40	5
2	4	5	3	1	41	6
2	5	1	3	4	42	4
2	5	1	4	3	43	5
2	5	3	1	4	44	5
2	5	3	4	1	45	6
2	5	4	1	3	46	6
2	5	4	3	1	47	7

3	1	2	4	5	ρ	$\nu(\rho)$
3	1	2	5	4	48	2
3	1	2	5	4	49	3
3	1	4	2	5	50	3
3	1	4	5	2	51	4
3	1	5	2	4	52	4
3	1	5	4	2	53	5
3	2	1	4	5	54	3
3	2	1	5	4	55	4
3	2	4	1	5	56	4
3	2	4	5	1	57	5
3	2	5	1	4	58	5
3	2	5	4	1	59	6
3	4	1	2	5	60	4
3	4	1	5	2	61	5
3	4	2	1	5	62	5
3	4	2	5	1	63	6
3	4	5	1	2	64	6
3	4	5	2	1	65	7
3	5	1	2	4	66	5
3	5	1	4	2	67	6
3	5	2	1	4	68	6
3	5	2	4	1	69	7
3	5	4	1	2	70	7
3	5	4	2	1	71	8

4	1	2	3	5	ρ	$\nu(\rho)$
4	1	2	5	3	72	3
4	1	3	2	5	73	4
4	1	3	5	2	74	4
4	1	5	3	2	75	5
4	1	5	2	3	76	5
4	1	5	3	2	77	6
4	2	1	3	5	78	4
4	2	1	5	3	79	5
4	2	3	1	5	80	5
4	2	3	5	1	81	6
4	2	5	1	3	82	6
4	2	5	3	1	83	7
4	3	1	2	5	84	5
4	3	1	5	2	85	6
4	3	2	1	5	86	6
4	3	2	5	1	87	7
4	3	5	1	2	88	7
4	3	5	2	1	89	8
4	5	1	2	3	90	6
4	5	1	3	2	91	7
4	5	2	1	3	92	7
4	5	2	3	1	93	8
4	5	3	1	2	94	8
4	5	3	2	1	95	9

5	1	2	3	4	ρ	$\nu(\rho)$
5	1	2	3	4	96	4
5	1	2	4	3	97	5
5	1	3	2	4	98	5
5	1	3	4	2	99	6
5	1	4	2	3	100	6
5	1	4	3	2	101	7
5	2	1	3	4	102	5
5	2	1	4	3	103	6
5	2	3	1	4	104	6
5	2	3	4	1	105	7
5	2	4	1	3	106	7
5	2	4	3	1	107	8
5	3	1	2	4	108	6
5	3	1	4	2	109	7
5	3	2	1	4	110	7
5	3	2	4	1	111	8
5	3	4	1	2	112	8
5	3	4	2	1	113	9
5	4	1	2	3	114	7
5	4	1	3	2	115	8
5	4	2	1	3	116	8
5	4	2	3	1	117	9
5	4	3	1	2	118	9
5	4	3	2	1	119	10


représentation graphique de l'application
ρ ⟶ ν(ρ)
ν(ρ)
10
9
8
7
6
5
4
3
2
1
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110 115 120
ρ

VII — *CONCLUSION*

Pour clore ce chapitre, répondons à la perplexité, que nous devinons chez le lecteur devant un problème de classification, concernant le choix de l'espace où travailler et du critère à adopter. Un tel choix peut parfois s'imposer compte tenu de la nature des données ou des contraintes du problème. De toute façon ce choix n'est pas indépendant de l'existence d'algorithmes permettant de définir un optimum (au moins relatif) de la quantité critère.

Nous plaiderons pour les méthodes les moins arbitraires et les plus générales quand, bien sûr, cette généralité ne compromet pas la fécondité des résultats qu'on obtient.

Les techniques qui se voudraient le moins arbitraires opéreraient directement à partir de la matrice d'incidence des données, sans mettre à l'entrée ni indice de similarité, ni ordonnance associée ; mais ces méthodes, comme on l'a vu [cf. § II.2.2 et 2.3], conservent toujours un certain degré d'arbitraire. D'autre part, il n'est pas sûr, à moins de se limiter à un tableau de taille modérée, qu'un algorithme défini dans ce cadre n'introduise finalement plus d'arbitraire que ne le ferait le choix d'un indice de similarité ou de la préordonnance associée. Enfin ces méthodes ne sont pas naturellement généralisables au cas de la recherche d'une hiérarchie de classifications.

Bien que présentant un certain arbitraire concernant le choix du seuil s [cf. III.1], la méthode où la donnée est une relation symétrique R sur E semble d'une grande généralité, mais ici, comme le montre le résultat de Moon, on a trop peu retenu d'information pour obtenir un résultat significatif : en effet, nous avons déjà fait remarquer que si $|R| < \frac{1}{2}\binom{n}{2}$ $\left(\text{resp. } |R| > \frac{1}{2}\binom{n}{2}\right)$ la relation d'équivalence Π relative à la partition la plus fine (resp. la moins fine) est, pourvu que n soit assez grand, une très bonne approximation de R au sens du critère $|R \Delta \Pi|$.

Quand les différentes partitions définies par les différents caractères a_h [cf. § IV. 1] peuvent être considérées comme également discriminantes, ou, ce qui revient au même, quand deux modalités quelconques relatives à un seul ou à deux caractères peuvent être regardées comme également significatives de la ressemblance, la méthode des partitions centrales de S. Régnier s'impose pour la recherche d'une partition, d'autant plus si on tient compte de la performance de l'algorithme attaché à cette méthode.

Toutefois l'imperfection des données ne permet pas, le plus souvent, les hypothèses dans lesquelles la méthode précédente s'inscrit. Les effets de cette imperfection se font beaucoup moins sentir lorsque la donnée de base retenue est une préordonnance sur E. [cf. Ch. 1 § III]. Il est satisfaisant de constater que, dans ce cadre, la recherche d'une partition apparaît comme un cas particulier de celui de la recherche d'une chaîne de partitions : la donnée d'une partition étant assimilée à celle d'une préordonnance ultramétrique à deux classes. Les études comparatives ainsi que certains aspects statistiques que nous avons développés nous permettent de contrôler ces méthodes beaucoup mieux que les précédentes.

Pour le cas de la recherche d'une partition, on adoptera la quantité critère, ((1) du § IV.A.5), obtenue en réduisant celle de La Véga. Par ailleurs, il existe un algorithme simple et naturel qui détermine une chaîne de partitions optimale au sens du critère d'ordre lexicographique. Un tel critère vient d'être jugé, (§ VI.B.3), du point de vue de celui, plus classique, du nombre d'inversions. Sur cette chaîne, il nous sera possible de détecter les niveaux les plus significatifs au moyen de la quantité critère (1) du paragraphe VI A.5.

CHAPITRE III

ALGORITHMES

I – *INTRODUCTION*

De manière intuitive, un algorithme est un procédé de calcul mécaniquement réalisable. Le but du calcul est ici de déterminer une classification (i.e. partition) ou une hiérarchie de classifications (i.e. chaîne de partitions) qui respecte de la manière la plus satisfaisante les ressemblances entre objets. Compte tenu de la forme que nous avons donnée au problème, la propriété la plus appréciable d'un algorithme est de déterminer, en un nombre de pas raisonnable, la partition ou la chaîne de partitions qui optimise un critère donné.

Les algorithmes exposés par Sokal et Sneath dans leur traité ne répondent pas à des critères tels que nous les avons définis. La plupart d'entre eux sont des procédures d'approximation, plus ou moins justifiées, qui restent parfois ambiguës. Néanmoins ces méthodes, selon les auteurs, donnent des résultats très voisins entre eux et ceux des méthodes usuelles (non automatiques). De toute façon, nous ne prétendons pas que les méthodes que nous allons présenter donneront, systématiquement, dans la pratique, de meilleurs résultats, (une classification doit répondre à tant d'exigences, certaines contradictoires !), cependant, ces méthodes sont mathématiquement bien définies et par conséquent leurs propriétés peuvent être dégagées de façon rigoureuse.

Nous commencerons par exposer un algorithme de recherche d'une classification qui permettra de cheminer le long d'une suite de partitions, remplaçant une partition par une autre plus avantageuse, (pour un critère donné). Cet algorithme dit des "transferts" a été développé par S. Régnier pour le calcul des partitions centrales [cf. Ch.2 § IV.3.1].

Notre souci d'établir une technique de recherche d'une hiérarchie de classifications qui soit très liée à une donnée de base, préordonnance sur E, nous a fait aboutir à un algorithme dont le caractère naturel est exprimé au moyen d'une propriété d'optimalité relativement au critère d'ordre lexicographique [cf. Ch.2 § VI A.5.1 et B.2.3]. Nous nous sommes rendus compte que notre algorithme était analogue à celui de Sneath (1957) ; il reste néanmoins plus systématique puisqu'il ne tient compte que de la préordonnance de départ. J.P. Benzécri soulignait la clarté de l'algorithme de Sneath mais regrettait que "seul un lien ténu pouvait par enchaînements successifs relier des points éloignés" ; c'est là une des raisons pour lesquelles nous avons tenu à juger la proximité lexicographique du point de vue du critère du nombre d'inversions [§ VI B.3]. De toute façon on se doute que la solution de l'algorithme sera d'autant meilleure que l'aptitude de la population E à être organisée en une hiérarchie de classifications qui respecte les ressemblances entre objets est plus grande ; nous préciserons ce point au chapitre suivant.

II – *ALGORITHME DES TRANSFERTS*

1 – Définition

P désigne une partition quelconque de E et $C(P)$ la valeur d'un critère C, qu'il s'agit, par exemple, de maximiser. A partir de P on peut construire une partition P' particulièrement voisine en effectuant le transfert d'un objet d'une classe dans une autre. En étiquetant les classes de la partition, on représentera P au moyen d'une application qui à chaque objet associe le numéro de la classe qui le contient

$$E \xrightarrow{f} \{1, 2, \ldots, k\}.$$

Le transfert de l'objet y de la classe l dans la classe m est défini à partir de l'application f, de la façon suivante :

$$f \longrightarrow f' : \begin{cases} f'(y) = m \neq f(y) = l \\ f'(x) = f(x) \quad \text{si} \quad x \in E,\ x \neq y \end{cases}$$

ce transfert est noté $T_l^m(y)/f$.

En répétant l'opération on peut parcourir tout l'ensemble $\mathcal{P}_k$ des partitions en k classes au plus. (Pour $k = n$, $\mathcal{P}_n$ est alors l'ensemble de toutes les partitions possibles, $\mathcal{P}$).

L'algorithme consiste à cheminer sur $\mathcal{P}_k$ par transfert, en choisissant chaque fois le transfert $T_l^m(y)/f$ qui augmente le plus la quantité critère $C(P)$: $C(P') - C(P)$ maximum.

On arrêtera le cheminement lorsqu'aucun transfert avantageux n'est possible ; en d'autres termes, P est la partition déterminée par l'algorithme si P est atteinte et si pour tout P' se déduisant de P par transfert : $C(P') - C(P) < 0$.

Relativement à la technique ci-dessus, il apparaît une nouvelle métrique τ : $\tau(P_1, P_2)$ = nombre minimum de transferts nécessaires pour passer d'une partition à l'autre (de P_1 à P_2). La partition obtenue par l'algorithme réalise un optimum local de la fonction $C(P)$ au sens de la métrique τ. Il peut exister plusieurs optimums locaux.

2 – Calcul des partitions centrales

Dans ce cas, relativement au cube des relations sur $E, \{0, 1\}^{n \times n}$, le problème est linéaire, puisqu'il s'agit de maximiser une quantité critère qui se présente comme la forme linéaire

$$L(P) = \sum_{(x,y)} t_{xy}\, \overline{\omega}_{xy} \quad \left(t_{xy} = s_{xy} - \frac{1}{2}\right)$$

où $\omega_{xy} = \omega_{yx}$ pour tout couple $(x, y) \in E \times E$

et $\omega_{x,z} \geqslant \omega_{xy}\, \omega_{yz}$ pour tout x, y et z de E .

[cf. Ch.2 § IV 3.2].

Nous avons déjà vu qu'on pouvait écrire

$$L(P) = \sum_{h=1}^{k} \left(\sum_{E_h \times E_h} t_{xy} \right) = \sum_{h} \sum_{f(x) = f(y) = h} t_{xy}$$

où f est l'application définissant l'étiquetage des classes de la partition.

Considérant la matrice

$$(\omega_{xy}),\ (\omega_{xy} = 1 \quad \text{si} \quad f(x) = f(y) \quad \text{et} \quad 0 \text{ sinon}),$$

le transfert ne modifie que la ligne et la colonne de l'objet y transféré :

ω_{xy} $(= \omega_{yx})$ passe de 1 à 0 si $f(x) = l$ et $x \neq y$

de 0 à 1 si $f(x) = m$.

La variation de L est donc

$$\Delta = -\sum_{f(x) = l} (t_{xy} + t_{yx}) + \sum_{f(x) = m} (t_{xy} + t_{yx})$$

En posant

$$C_x^l = \sum_{\substack{f(y) = l \\ x \neq y}} t_{xy}$$

= "attraction de l'objet x par la classe l" ,

alors $$\Delta = 2(C_y^m - C_y^l)$$

donne le gain du transfert de l'objet y dans la classe m ; et le choix du transfert le plus avantageux repose seulement sur la connaissance de la matrice des C_x^l (n lignes × k colonnes).

Au cours du transfert $T_l^m(y)/f$, cette matrice se modifie selon des règles simples

$$\begin{cases} \text{les colonnes } l \text{ ou } m \text{ sont seules modifiées,} \\ C_x^l \text{ devient } C_x^l - t_{xy}\,, \\ C_x^m \text{ devient } C_x^m + t_{xy}\,. \end{cases}$$

Exemple : considérons la partition $\{\{a,d,e,f\},\{b,c\}\}$ de l'ensemble $E = \{a,b,c,d,e,f\}$, dans l'exemple (Ch.2 § IV.3.2), où nous attribuerons le n° 1 à la classe $\{a,d,e,f\}$ et le n° 2 à $\{b,c\}$. Nous avons pu constater que le transfert de l'objet e à la classe de n° 2 était avantageux.

Remarque

Il importe de noter qu'en vertu de l'argument présenté au paragraphe IV.4. du chapitre 2, deux objets équivalents sont classés ensemble par l'algorithme.

3 — Application de l'algorithme à d'autres critères

L'algorithme des transferts a été utilisé avec succès pour le critère de La Véga qui, rappelons-le, s'exprime de la façon suivante :

$$C(P) = \sum_{F\times F} I_{R\times S}(p,q)\left[I_\omega(p,q) - \frac{1}{2}\right]$$

[cf. Ch.2 § VI A.2.2 formule (2)].

Calculons la variation de $C(P)$ pour le transfert de l'objet e de la classe E_l, à laquelle il appartient, dans E_m. Pour cela désignons par

$$A_e = \bigcup_{\substack{x\in E_l \\ x\neq e}} \{x,e\} \text{ et par } B_e = \bigcup_{y\in E_m} \{y,e\}.$$

Ce transfert définit les transformations

$$R \longrightarrow R - A_e + B_e$$

$$S \longrightarrow S - B_e + A_e$$

d'où

$$R\times S \longrightarrow ((R-A_e)+B_e)\times((S-B_e)+A_e)$$

(addition et soustraction ensemblistes).

Calcul exécuté, la variation de $C(P)$ est définie par :

$$\Delta(C(P)) = \sum_{(p,q)\in F\times F} [I\ (p,q) - - I_L(p,q)]\left[I_\omega(p,q) - \frac{1}{2}\right]$$

où

$$K = (R - A_e)\times A_e + B_e \times (S - B_e) + B_e \times A_e$$

et

$$L = (R - A_e)\times B_e + A_e \times (S - B_e) + A_e \times B_e$$

(somme et différence ensemblistes).

Il serait encore intéressant d'appliquer cet algorithme avec le critère que nous proposons

$$(|\omega \cap R\times S| - r.s/2)/\sqrt{r.s} \quad \text{[cf. Ch.2 § VI A.5.]}$$

III — *ALGORITHME "LEXICOGRAPHIQUE"*

C'est ainsi que nous appellerons l'algorithme que nous avons annoncé dans l'introduction, de recherche d'une hiérarchie de classifications. Compte tenu de la forme que nous avons donnée au problème, [cf. Ch.2 — § B-1.1 et 1-3] nous commencerons par définir une préordonnance ultramétrique ω_u sur E, liée à ω au moyen d'une fonction ordinale H sur F pour laquelle :

$$H(x,y) \leqslant r \text{ et } H(y,z) \leqslant r \implies H(x,z) \leqslant r$$

pour tout x, y et z de E. Nous exposerons ensuite l'algorithme, qui détermine ce préordre total sur F par ses sections commençantes, (qui sont des parties saturées de F), et démontrerons son caractère optimal.

Nous terminerons en nous intéressant à l'étude de l'algorithme dans le cas où la donnée est une ordonnance sur E (ordre total sur F).

1 – Définition d'une préordonnance ultramétrique ω_u liée à ω

ω est la préordonnance donnée sur E et $\rho(p)$ désigne le rang pour d'un élément p quelconque de F ; [cf. Ch.2 § VI A.5.1]. $c(x, y)$ indiquera une chaîne finie d'éléments de E d'origine x et d'extrémité y telle que $(z_1, z_2, \ldots, z_k)$; $z_1 = x$, $z_k = y$ et $z_i \neq z_j$ pour tout $i \neq j$. $\mathcal{C}(x, y)$ sera l'ensemble de toutes les chaînes d'origine x et d'extrémité y.

Enfin $H(x, y)$ désignera une fonction ordinale sur F compatible avec la préordonnance ultramétrique ω_u à déterminer, c'est-à-dire telle que :

$$H(p) \leqslant H(q) \Longleftrightarrow p \leqslant q$$

pour ω_u ; pour tout p et q de F.

1.1 – *Définition de $H(x, y)$*

Les opérations de maximum et de minimum dans N étant représentées par $\vee$ et $\wedge$, nous définissons

$$\varphi(c(x, y)) =$$
$$= \rho(z_1, z_2) \vee \rho(z_2, z_3) \vee \ldots \vee \rho(z_{(k-1)}, z_k)$$

où $c(x, y) = (z_1, z_2, \ldots, z_k)$.

On posera

$$(1)\quad H(x, y) = \bigwedge_{c \in \mathcal{C}(x, y)} \varphi(c(x, y))$$

$H(x, y) = \rho(x, y)$ si le minimum de $\varphi(c(x, y))$ est atteint pour la chaîne réduite à (x, y).

Nous allons montrer que $H(x, y)$ définit bien une préordonnance ultramétrique, c'est-à-dire

1.2 – *Proposition*

$$H(x, y) \leqslant r \text{ et } H(y, z) \leqslant r \Longrightarrow H(x, z) \leqslant r$$

pour tout x, y et z de E.

$H(x, y) \leqslant r$ signifie qu'il existe une chaîne $c(x, y)$ pour laquelle $\varphi(c(x, y)) \leqslant r$ soit par exemple

$$c(x, y) = (u_1, u_2, \ldots, u_k)$$

$u_1 = x$, $u_k = y$ et $u_i \neq u_j$ pour $i \neq j$.

$H(y, z) \leqslant r$ signifie qu'il existe une chaîne $c(y, z)$ pour laquelle $\varphi(c(y, z)) \leqslant r$ soit par exemple

$$c(y, z) = (v_1, v_2, \ldots, v_h)$$

$v_1 = y$, $v_h = z$ et $v_i \neq v_j$ pour $i \neq j$.

Désignons par u_l le premier élément de la suite $c(x, y)$ qui se retrouve dans la suite $c(y, z)$; cet élément existe nécessairement puisque les deux suites ont en commun y.

Soit $u_l = v_m$. Dans ces conditions la suite :

$$(u_1, u_2, \ldots, u_l, v_{m+1}, v_{m+2}, \ldots, v_h)$$

est un élément $c(x, z)$ de $\mathcal{C}(x, z)$. En vertu de la définition de φ et des hypothèses, on a pour ce $c(x, z)$, $\varphi(c(x, z)) \leqslant r$.

Donc à fortiori $H(x, z) \leqslant r$.

2 – Algorithme permettant d'obtenir la préordonnance ultramétrique ω_u définie par H

Soient $B_1, B_2, \ldots, B_p$, les différentes classes du préordre total ω, avec $B_1 < B_2 < \cdots < B_p$ pour l'ordre total quotient et soit $C_1, C_2, \ldots, C_p$, la suite obtenue de la façon suivante :

$$C_1 = B_1,\ C_2 = B_1 \cup B_2, \ldots, C_p = B_1 \cup B_2 \cup \ldots \cup B_p$$

Pour tout i, C_i est une section commençante de F pour ω.

Notons $\overline{C}_i$ la saturée de C_i, (rappelons que si A est une partie de F, la saturée $\overline{A}$ est l'ensemble des paires, $\{x, y\}$, telles que : $\{x, y\} \in \overline{A}$ si $\{x, y\} \in A$ ou $(\exists z) : \{x, z\} \in A$ et $\{y, z\} \in A$).

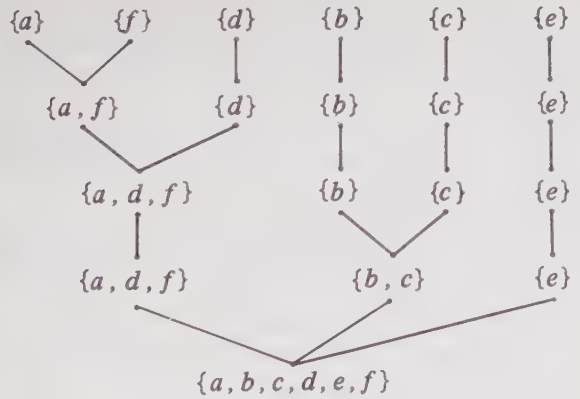
Soit q le plus grand indice pour lequel la saturée $\overline{C}_q$ est identique à F ($\overline{C}_q = F$) ; la suite des saturées : $\overline{C}_1, \overline{C}_2, \ldots, \overline{C}_q$ est une suite croissante au sens de l'inclusion, puisque $C_i \subset C_j$ pour $i < j$. Cette suite est, précisément, celle des différentes sections commençantes de la préordonnance ultramétrique ω_u définie par $H(x, y)$.

Exemples : Nous nous proposons d'appliquer l'algorithme pour le tableau d'incidence (Ch.1 § II.2) sur la base de chacune des préordonnances : $\underline{\omega}$ et $\overline{\omega}$ (Ch.1 § III.1), ω^1 (la préordonnance "la plus fine", Ch.1 § IV.1), ω^2 (la préordonnance "la moins fine", Ch.1 § IV.2).

1′ – Application de l'algorithme à $\underline{\omega}$

af	ad	df	bc	ac	ef	ab	bd	cd	be	ce	de	bf	cf	ae
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	2	3	4	4	4	4	4	4	4	4	4	4	4	4

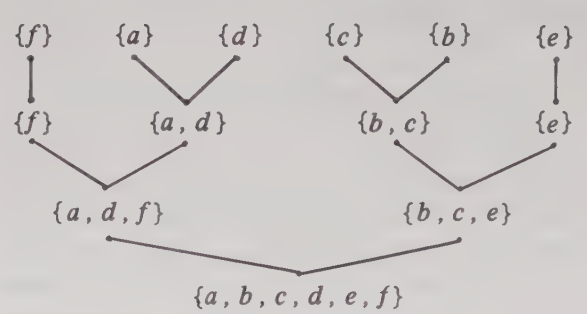
Le vecteur de composantes 1 ou 0 du niveau i représente la fonction caractéristique de $\overline{C}_i$. Le vecteur de la dernière ligne représente la fonction ordinale qui définit la préordonnance ultramétrique ω_u associée à $\underline{\omega}$. Une quelconque des composantes de ce vecteur est obtenue en retranchant à 5 (nombre maximum de niveaux – 1) le nombre de 1 de la colonne correspondante. La chaîne de partitions associée est la suivante :



2′ – Application de l'algorithme à $\overline{\omega}$

ad	bc	be	af	df	de	bd	ce	ef	ab	ac	cd	ae	bf	cf
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	0	0	1	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	3	3	3	4	4	3	4	4	4	4	4	4	4

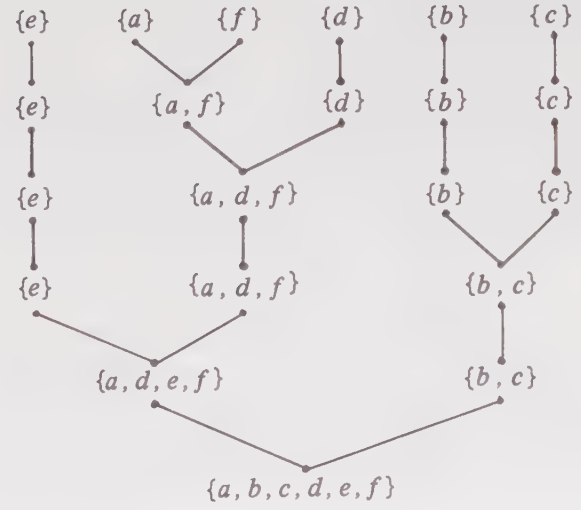
La chaîne de partitions associée est la suivante :



3′ – Application de l'algorithme à la préordonnance "la plus fine"

af	ad	df	bc	ef	ac	be	de	ce	bd	ab	cd	bf	ae	cf
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	0	0	1	0	0	0	0	0	1	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0	1	1	2	3	4	4	3	4	4	4	4	4	3	4

La chaîne de partitions associée est la suivante :

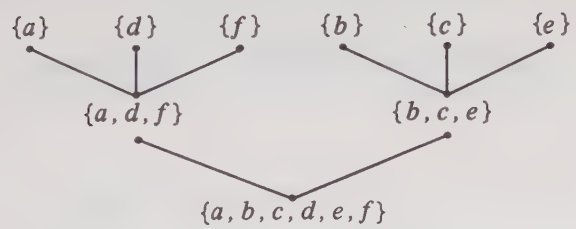


La préordonnance "la plus fine" étant une ordonnance, la chaîne de partitions associée est élémentaire [cf. § ci-dessus].

4′ – Application de l'algorithme à la préordonnance "la moins fine"

$ad = af = bc = be = df < ab = ac = ae = bd = bf = cd = ce = de = ef < cf$

1	1	1	1	1	0	0	0	0	0	0	1	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	3	3	3	3	4	4	4	4	4	4	3	4	4	4



2.3 – *Théorème* (Propriété optimale de la préordonnance ultramétrique ω_u).

Parmi les préordonnances ultramétriques de même type que ω_u , ω_u est la plus voisine de ω au sens du critère d'ordre lexicographique, [cf. Ch.2 § VI A.5.1 (a)]. De plus toute préordonnance ultramétrique dont la suite des sections commençantes,

$$R_1 , R_2 , \ldots , R_l , \ldots ,$$

satisfait la condition :

$$C_1 \subset R_1 , C_2 \subset R_2 , \ldots , C_l \subset R_l , \ldots ,$$

est telle que $\bar{C}_1 \subset R_1 , \bar{C}_2 \subset R_2 , \ldots , \bar{C}_l \subset R_l , \ldots ,$

Preuve

Soit $t = (t_1 , t_2 , \ldots , t_q)$ le type de ω_u ; on a :

$$\sum_{i=1}^{l} t_i = |\bar{C}_l| \geqslant |C_l|$$

Désignons par $(T_1 , T_2 , \ldots , T_q)$ la suite des sections commençantes d'un préordre ultramétrique, quelconque, sur F, de type t. Parmi les préordonnances ultramétriques de type t, toute celle pour laquelle $C_1 \subset T_1$ est plus proche de ω que de toute celle pour laquelle ce n'est pas le cas. D'autre part si $C_1 \subset T_1$, alors $\bar{C}_1 \subset T_1$ donc $\bar{C}_1 = T_1$ puisque $|\bar{C}_1| = |T_1|$.

De nouveau, parmi les préordonnances ultramétriques de type t, pour lesquelles $T_1 = \bar{C}_1$, toute celle pour laquelle $C_2 \subset T_2$ est plus proche de ω que toute celle pour laquelle ce n'est pas le cas. D'autre part si $C_2 \subset T_2$, alors $\bar{C}_2 \subset T_2$, donc $\bar{C}_2 = T_2$ puisque $|\bar{C}_2| = |T_2|$ et ainsi de suite . . . d'où le résultat annoncé.

La deuxième partie de la proposition est évidente, elle exprime une propriété de "finesse" de ω_u. $\bar{C}_1$ est la première section commençante, la plus petite (au sens de la cardinalité), pour une préordonnance ultramétrique dont la première section commençante contient C_1 ; $\bar{C}_2$ est la deuxième section commençante la plus petite pour une préordonnance ultramétrique dont la première section est $\bar{C}_1$ et dont la seconde contient C_2 ; etc.

Il résulte de la définition de l'algorithme la

2.4 – *Proposition*

La condition nécessaire et suffisante pour que la préordonnance ultramétrique ω_u, associée à ω par l'algorithme, soit comparable[1] avec ω est que la saturée de toute section strictement commençante (pour ω), C_i, soit une section commençante, au sens large, (pour ω).

Nous voulons dire que pour tout i, il existe h, tel que :

$$C_i = C_{i+h} \cup A_{i+h+1}$$

où A_{i+h+1} est une partie de B_{i+h+1}.

Nous allons à présent dégager certaines propirétés de l'algorithme dans le cas où la donnée ω est une ordonnance sur E, (ordre total sur F).

---

(1) Si ω et ω' sont deux préordres totaux sur F, ω et ω' sont dits comparables si le graphe de l'un est contenu dans celui de l'autre. Ici ω_u est moins fine que ω .

3 – Hiérarchie de classifications associée à une ordonnance

3.1 – *Proposition*

Dans le cas où la donnée ω est une ordonnance, la chaîne de partitions, définie par l'algorithme, est élémentaire [cf. Ch. 0 § I.2].

Ce cas est caractérisé par la propriété : chacune des classes B_i, (ci-dessus définies), se réduit à un élément de F ; on a dans ces conditions $f = n(n-1)/2$ classes. Comparons $\overline{C}_{i+1}$ et $\overline{C}_i$.

On a

$$\overline{\overline{C}_{i+1}} = \overline{\overline{C}_i \, B_{i+1}}$$

soit, par exemple, $\{x, y\}$ l'élément unique de la classe B_{i+1}. Deux cas sont possibles :

a) $\{x, y\} \in \overline{C}_i$; et alors, $\overline{C}_{i+1} = \overline{C}_i$,

b) $\{x, y\} \notin \overline{C}_i$; dans ces conditions, la partition définie par $\overline{C}_{i+1}$ se déduit de celle définie par $\overline{C}_i$, par réunion d'exactement deux classes, celle de x et celle de y.

D'après la proposition 2.4 ci-dessus, ω est une ordonnance pour laquelle la préordonnance ultramétrique, ω_u, associée par l'algorithme est comparable avec ω, si et seulement si la saturée de toute section, C, commençante pour ω, est une section commençante pour ω.

Une question se pose que nous allons exprimer de manière intuitive : "La propriété, qu'on vient d'énoncer, est-elle rare ?" Nous sommes ainsi amenés à étudier, dans l'ensemble, Ω, de toutes les ordonnances sur E, la proportion de celles, ω_u, pour lesquelles l'algorithme définit une préordonnance ultramétrique, ω_u, comparable avec ω ; nous nous rendrons compte, en particulier, du comportement de cette proportion lorsque n augmente, $(n = |E|)$.

Le cardinal de l'ensemble, Ω, de toutes les ordonnances est f ! soit

$$(n(n-1)/2)! \, , f = |F| = n(n-1)/2 \, .$$

Désignons par $\Omega_u(E)$ l'ensemble des ordonnances pour lesquelles la propriété ci-dessus est satisfaite. Le cardinal de $\Omega_u(E)$ dépend seulement de l'entier n, cardinal de E, posons, dans ces conditions, $\gamma(n) = |\Omega_u(E)|$. Le rapport qui nous intéresse est $\gamma(n)/f$!

3.2 – *Dénombrement de* Ω_u

Une ordonnance ω est un élément de Ω_u si et seulement si ω est compatible avec une préordonnance ultramétrique ω_u, associée à une chaîne élémentaire de partitions, [cf. Prop. 3-1].

Si ω_u est une telle préordonnance, qu'on dira élémentaire, ω_u comporte $(n-1)$ classes ; en désignant, respectivement par $t_1, t_2, \ldots, t_{(n-1)}$ les cardinaux des différentes classes (du préordre), prises de gauche à droite $(t_1, t_2, \ldots, t_{(n-1)})$ sera le type de ω_u. Il y a $t_1 ! \, t_2 ! \ldots t_{(n-1)}!$ ordonnances de Ω_u compatibles avec ω. Un calcul direct de $\gamma(n)$ suppose la détermination de l'ensemble des types des préordonnances ultramétriques élémentaires. Plutôt que de procéder directement, nous nous contenterons d'établir une relation de récurrence pour $\gamma(n)$.

3.2.1 – *Proposition*

On a, pour la fonction $\gamma(n)$, la relation de récurrence :

$$\gamma(n) = \sum_{1 \leqslant k \leqslant \frac{n}{2}} \gamma(k)\, \gamma(n-k)\, (k(n-k))\, ! \binom{n}{k} \binom{n-2}{k-1}$$

Le niveau $(n-1)$ de la chaîne élémentaire de partitions correspond à une partition de E en deux classes. Le type de partition du niveau $(n-1)$ est donc de la forme $(n-k, k)$ où k peut prendre les valeurs : $1, 2, \ldots, [n/2]$; (rappelons que $[n/2]$ est la partie entière de $n/2$).

Il y a $\binom{n}{k}$ partitions de E en deux classes de type $(n-k, k)$.

Considérons la partition de Ω_u selon le type de la partition obtenue par l'algorithme au niveau $(n-1)$ et désignons par Ω_k le sous-ensemble de

Ω_u pour lequel la partition obtenue au niveau $(n-1)$ est de type $(n-k, k)$ en posant $\gamma(n, k)$ le cardinal de Ω_k, on a

$$\gamma(n) = \sum_{1 \leqslant k \leqslant \frac{n}{2}} \gamma(n, k) . \qquad (1)$$

Considérons une partition particulière de type $(n-k, k)$ que nous noterons $(\overline{n-k}, \overline{k})$, soit $\delta(n, k)$ le nombre d'ordonnances pour lesquelles la partition obtenue par l'algorithme au niveau $(n-1)$ est précisément $(\overline{n-k}, \overline{k})$. Par raison de symétrie, $\delta(n, k)$ est le même quelle que soit la partition de type $(n-k, k)$. Donc :

$$\gamma(n, k) = \binom{n}{k} \delta(n, k)$$

et

$$\gamma(n) = \sum_{1 \leqslant k \leqslant \frac{n}{2}} \binom{n}{k} \delta(n, k) . \qquad (2)$$

Désignons, respectivement, par L et par M les ensembles des parties à deux éléments, de $(\overline{n-k})$ et de $\overline{k}$. Si ω est l'ordre total sur F, ω_1 et ω_2 seraient respectivement les restrictions de ω à L et à M, ω_{1u} et ω_{2u} les préordonnances ultramétriques associées à ω_1 et à ω_2 par l'algorithme.

La condition nécessaire et suffisante pour qu'une ordonnance ω sur E appartienne à Ω_u et que de plus le niveau $(n-1)$ de la chaîne de partitions associée par l'algorithme corresponde à la partition $(\overline{n-k}, \overline{k})$ est que :

a) $\omega_1 \in \Omega_u(\overline{n-k})$ et $\omega_2 \in \Omega_u(\overline{k})$,

b) LUM définisse une section de F commençante pour ω, et :

c) chaque classe de ω_{1u} ou ω_{2u} définisse un intervalle de F pour ω ; les différents intervalles, de la suite des intervalles que définit ω_{1u}, s'intercalant avec les différents intervalles de la suite des intervalles que définit ω_{2u}.

Rappelons que ω_{1u} et ω_{2u} comportent respectivement $(n-k-1)$ classes et $(k-1)$ classes.

$\delta(n, k)$ est le nombre de façons dont on peut constituer une ordonnance, sur E, remplissant les conditions *a*, *b* et *c*.

Pour définir ω, il y a lieu :

1/ de déterminer une ordonnance ω_1 sur $(\overline{n-k})$ appartenant à $\Omega_u(n-k)$, cela peut se faire de $\gamma(n-k)$ façons différentes ;

2/ de déterminer une ordonnance ω_2 sur $\overline{k}$ appartenant à $\Omega_u(k)$, cela peut se faire de $\gamma(k)$ façons différentes ;

3/ d'intercaler les différents intervalles de la suite des intervalles définie par ω_{1u} avec les différents intervalles définis par ω_{2u}. Cela peut se faire de $\binom{n-2}{k-1}$ façons différentes, puisque ω_{1u} comporte $(n-k-1)$ classes et ω_{2u} $(k-1)$ classes.

4) enfin de définir un ordre total sur les $k(n-k)$ paires $\{x, y\}$ pour lesquelles x appartient à $\overline{k}$ et y à $(\overline{n-k})$, cela peut se faire de $(k(n-k))!$ manières différentes.

On a donc bien

$$\delta(n, k) = \gamma(n-k)\,\gamma(k) \binom{n-2}{k-1} (k(n-k))!$$

3.2.2 – *Remarques*

$\gamma(n)$ augmente très vite lorsque n croît, le rapport $\gamma(n)/f!$ $(f = n(n-1)/2)$ tend très rapidement vers 0 lorsque n augmente pour tendre vers l'infini.

n	3	4	5	6	7	8	9	10	11	12	13	14	15
$\gamma(n) \geqslant$	6	10^2	10^5	10^{10}	10^{14}	10^{21}	10^{30}	10^{40}	10^{51}	10^{65}	10^{81}	10^{99}	10^{118}
$\frac{\gamma(n)}{f!} \leqslant$	1	6/10	$5/10^2$	$3/10^3$	$1/10^5$	$3/10^8$	$3/10^{12}$	$2/10^{16}$	$8/10^{22}$	$2/10^{27}$	$2/10^{34}$	$9/10^{42}$	$3/10^{50}$

4 – Extension de l'algorithme "lexicographique" pour la recherche d'une famille de suites de recouvrements

Rappelons qu'un recouvrement d'un ensemble E est défini par la donnée d'une famille de parties de E, dont la réunion est E ; une partition étant un recouvrement particulier. A un recouvrement Q est associée la relation binaire :

$(\forall x, y \in E) : xQy \Longleftrightarrow x$ et y appartiennent à une même partie. Un recouvrement Q sera dit plus fin qu'un recouvrement Q' si $xQy \Longrightarrow xQ'y$ pour tout x et y de E.

Soit $Q^1, Q^2, \ldots, Q^h, \ldots$ une suite de recouvrements de E, totalement ordonnée par finesse décroissante ; une telle suite respecte les ressemblances entre objets si le premier indice h, pour lequel deux objets donnés quelconques de E se trouvent réunis dans une même partie de Q^h, est d'autant plus petit que la ressemblance des deux objets est grande.

L'algorithme que nous définirons nous permettra d'obtenir, de manière naturelle, une famille finie de suites totalement ordonnées de recouvrements dont chacune respecte de manière satisfaisante, pour un degré de finesse donné, les ressemblances entre objets. Ce point sera clarifié par ce qui suit.

Reprenons les notations du paragraphe 2 précédent ; $B_1, B_2, \ldots, B_p$ désignent les différentes classes du préordre ω, avec $B_1 < B_2 < \cdots < B_p$ pour l'ordre total quotient ; $C_1, C_2, \ldots, C_p$ est la suite des sections strictement commençantes de F pour ω, soit

$$C_i = B_1 \cup B_2 \cup \ldots \cup B_i$$

soit d'autre part, pour $i < p$ et pour $h \leqslant p - i$, l'intervalle

$$D_i^h = B_{i+1} \cup B_{i+2} \cup \ldots \cup B_{i+h}$$

soit enfin $(P_0, P_1, \ldots, P_q)$ la suite des partitions définie par l'algorithme "lexicographique" ; P_0 est la partition pour laquelle chaque classe contient un seul objet et P_i celle déterminée par la saturée $\bar{C}_{i-1}$ de C_{i-1}.

A chaque partition P_i est associée l'une des suites de recouvrements (Q_i^h), $h = 1, 2, \ldots, (p - i)$, de la façon suivante :

Notons E_l une quelconque des classes de P_l. Pour h fixé la saturée $\bar{D}_{i-1}^h$ de D_{i-1}^h définit une partition sur une partie de E ; e_m désignera une classe quelconque de cette partition. Une partie de E est un élément du recouvrement Q_i^h cherché si et seulement si cette partie est une classe de type E_l ou une classe de type e_m. On remarquera que la suite (Q_0^h) n'est pas autre que la suite

$$(P_0, P_1, P_2, \ldots, P_q).$$

Dans la pratique on est surtout intéressé par des recouvrements faiblement empiétants de sorte que nous proposons de déterminer en même temps que la suite $(P_0, P_1, \ldots, P_q)$, la suite

$$(Q_1^1, Q_2^1, \ldots, Q_{q-1}^1) ;$$

Q_l^1 étant défini à partir de P_l et de B_l.

Un tel algorithme est moins vulnérable à l'effet des "enchaînements successifs" auquel nous avons fait allusion dans l'introduction.

Exemple : Reprenons l'exemple (2′) du paragraphe 2 ci-dessus. La suite des recouvrements

$$(Q_1^1, Q_2^1, \ldots, Q_{q-1}^1)$$

est ici

$$Q_1^1 = \{\{f\}, \{a, d, f\}, \{b, c\}, \{b, e\}\}$$

$$Q_2^1 = \{\{a, d, f\}, \{b, c, e\}, \{d, e\}\}$$

q étant égal à 3.

CHAPITRE IV

ASPECTS STATISTIQUES

I – *INTRODUCTION*

La population E d'objets sur laquelle il s'agit de découvrir une classification ou une hiérarchie de classifications est en général un échantillon d'une population E' plus vaste, qu'il est souvent impossible d'observer dans sa totalité. Par ailleurs, une description exhaustive des éléments de E est encore souvent impossible ; les caractères ou attributs, retenus pour la description de E, constituent un échantillon d'un ensemble plus vaste de caractères ou attributs.

Sous ce point de vue, il est intéressant d'étudier certains problèmes statistiques, de nature asymptotique, dans le cadre d'hypothèses probabilistes.

Certaines distributions statistiques ont bien été étudiées dans le chapitre des critères, mais elles restaient intimement liées à la définition d'un critère et supposaient en outre fixés, un ensemble E d'objets et une ordonnance ω correspondante.

Nous commencerons par considérer deux problèmes de convergence statistique, inhérents à la méthode des partitions centrales [cf. Ch. 2 § IV] et développés par S. Régnier dans [26 (c)] que nous reprenons en simplifiant la présentation. Nous poserons ensuite un problème de test statistique que nous étudierons dans le cas où la donnée est une préordonnance sur E. Nous séparerons complètement ces deux études dont les buts sont distincts et qui se réfèrent à des données de base de nature différente.

II – *DEUX PROBLEMES DE CONVERGENCE RELATIFS A LA METHODE DES PARTITIONS CENTRALES*

A – PREMIER PROBLEME

1 – Introduction

Ici nous supposons que $E = \{x^1, x^2, \ldots, x^n\}$ est un échantillon de n éléments aléatoires indépendants pris dans un ensemble E' muni d'une mesure de probabilité Q'. E' peut éventuellement être infini ; en général, E' est fini de cardinal N assez grand et l'échantillon E peut être considéré comme obtenu au moyen d'un tirage avec remise dans E' ; de la sorte, la probabilité d'obtenir plus d'une fois le même objet n'est pas nulle, elle est petite si n est petit devant N.

L'application $a = (a_1, a_2, \ldots, a_p)$ qui définit la description de E' dans le produit cartésien

$$\mathfrak{A} = A_1 \times A_2 \times \ldots \times A_p,$$

où $\mathfrak{A}$ est *fini*, est supposée ici fixée. [cf. Ch.2 § IV.1] a sera supposée mesurable pour Q' et pour la mesure uniforme sur $\mathfrak{A}$ fini. Dans ces conditions, Q' induit sur $\mathfrak{A}$ une mesure de probabilité Q, définie comme suit : Si ν est un élément quelconque de $\mathfrak{A}$:

$$Q(\nu) = Q'\{a^{-1}(\nu)\} \qquad (1)$$

où $a^{-1}(\nu)$ est l'ensemble des éléments de E' dont l'image est ν.

L'application a définit encore une représentation de E dans $\mathfrak{A}$; les images $\nu_i = a(x^i)$, $i = 1, 2, \ldots, n$;

sont n éléments aléatoires indépendants de l'ensemble fini $\mathcal{A}$ muni de la mesure Q. $v_1, v_2, \ldots, v_n$ définissant sur $\mathcal{A}$ une mesure de probabilité empirique Q_n définie comme suit :

$$Q_n(v) = \frac{n_v}{n} \tag{2}$$

où v est un élément quelconque de $\mathcal{A}$ et n_v le nombre d'objets de E qui ont la même image v par a ; $n_v = |a^{-1}(v) \cap E|$.

Nous avons déjà vu, (proposition 4.1 § IV Ch.3), qu'une partition centrale sur E (resp. sur E') est image réciproque par a d'une partition centrale sur $\mathcal{A}$, lorsqu'on adopte sur $\mathcal{A}$ la mesure de probabilité Q_n (resp. Q). Nous travaillerons, dans ces conditions, sur $\mathcal{A}$ muni de la mesure de probabilité Q_n ou Q. $\mathfrak{X}$ désignera l'ensemble des partitions sur $\mathcal{A}$.

Soit Y une partition quelconque sur $\mathcal{A}$, ($Y \in \mathfrak{X}$) ; posons, conformément à la formule (1) du paragraphe IV.4 (Ch.2).

$$D_n(Y) = \sum_{(v,w) \in \mathcal{A} \times \mathcal{A}} (\eta_{vw} - S_{vw})^2 \, Q_n(v) \, Q_n(w) \tag{3}$$

où η_{vw} est la fonction caractéristique associée à la partition Y, ($\eta_{vw} = 1$ si v et w sont dans la même classe et 0 sinon) et $S_{vw} = \sum_{i=1}^{p} S_{vw}^h / p$, où S_{vw} est la fonction caractéristique de la partition sur $\mathcal{A}$ associée au caractère a_h.

Le centre C_n, ensemble des partitions centrales sur $\mathcal{A}$ muni de la mesure Q_n, est défini comme étant l'ensemble des éléments de $\mathfrak{X}$ où $D_n(Y)$ atteint sa borne inférieure :

$$C_n = \{Z : D_n(Z) \leqslant D_n(Y) \text{ pour tout } Y \text{ de } \mathfrak{X}\}.$$

D'une manière analogue, en posant

$$D(Y) = \sum_{(v,w) \in \mathcal{A} \times \mathcal{A}} (\eta_{vw} - S_{vw})^2 \, Q(v) \, Q(w)$$

étant muni de la mesure Q, le centre C est défini de la façon suivante :

$$C = \{Z : D(Z) \leqslant D(Y) \text{ pour tout } Y \text{ de } \mathfrak{X}\}.$$

Pour Y fixé, $D_n(Y)$ est une variable aléatoire définie sur $\mathcal{A}^n$ muni de l'ensemble de ses parties et de la mesure de probabilité induite par Q. Au vecteur aléatoire $(D_n(Y) : Y \in \mathfrak{X})$ est attaché le centre C_n qui est donc une partie aléatoire de $\mathfrak{X}$. Nous allons étudier le comportement asymptotique de C_n, $(n \longrightarrow \infty)$; certains des événements que l'on considérera se situent dans $\prod_{i=1}^{\infty} \mathcal{A}_i$ muni de la σ-algèbre produit $\bigotimes_{i=1}^{\infty} P(\mathcal{A}_i)$ [1] où $\mathcal{A}_i = \mathcal{A}$ pour tout i et où $P(\mathcal{A}_i)$ désigne l'ensemble des parties de $\mathcal{A}$.

2 – Proposition

On a presque sûrement : lim sup $C_n \subset C$, c'est-à-dire :

$$\Pr\{\limsup C_n \subset C = 1\} \tag{1}$$

d'autre part $$\lim_{n \to \infty} \Pr\{C_n \subset C\} = 1 \, . \tag{2}$$

Démonstration

L'événement complémentaire de lim sup $C_n \subset C$ peut être exprimé de la façon suivante :

$$(\exists X) : X \notin C \quad \text{et} \quad X \in \limsup C_n \, . \tag{3}$$

Pour X fixé n'appartenant pas à C, déterminons la probabilité de l'événement $X \in \limsup C_n$ qui signifie : la suite des C_n qui contiennent X est infinie, soit :

$$(\forall k)\ (\exists n > k) : \inf_{Y \in \mathfrak{X}} (D_n(Y) - D_n(X)) \geqslant 0 \, ;$$

un tel événement implique

$$\limsup_n \inf_{Y \in \mathfrak{X}} (D_n(Y) - D_n(X)) \geqslant 0 \tag{4}$$

d'où

$$\Pr\{X \in \limsup C_n\} \leqslant$$

$$\leqslant \Pr\{\limsup_n \inf_{Y \in \mathfrak{X}} (D_n(Y) - D_n(X)) \geqslant 0\} \, . \tag{5}$$

---

(1) Pour la définition d'une telle σ-algèbre on consultera par exemple : J. Neveu. *Bases Mathématiques du Calcul des Probabilités*, p. 74 à 79. Paris, Masson.

D'autre part, avec les notations de ci-dessus

$(\forall \nu \in \mathfrak{A})$: $Q_n(\nu)$ converge presque sûrement vers $Q(\nu)$ et donc

$(\forall Y \in \mathfrak{X})$: $D_n(Y)$ converge presque sûrement vers $D(Y)$.

Il en résulte, puisque $\mathfrak{X}$ est fini, que

$$\inf_{Y \in \mathfrak{X}} (D_n(Y) - D_n(X)) \xrightarrow[\text{p.s.}]{c} \inf_{Y \in \mathfrak{X}} (D(Y) - D(X))$$

Puisque X n'appartient pas à C, cette limite presque sûre est strictement négative et le second membre de (5) est nul. Donc pour X fixé dans $\mathfrak{X} - C$, $\Pr\{X \in \limsup C_n\} = 0$.

La probabilité de l'événement (3) est inférieure ou égale à :

$$\sum_{X \in \mathfrak{X} - C} \Pr\{X \in \limsup C_n\} = 0 \text{ (somme finie) ;}$$

ainsi se trouve démontrée (1).

La convergence presque sûre de $D_n(Y)$ vers $D(Y)$ pour tout Y de $\mathfrak{X}$ entraîne, celle, presque sûre de :

$$\inf_{Y \in \mathfrak{X}} D_n(Y) \text{ vers } \inf_{Y \in \mathfrak{X}} D(Y)$$

Il en résulte que $D_n(Y)$ (resp. $\inf_{Y \in \mathfrak{X}} D_n(Y)$), converge en probabilité vers $D(Y)$ (resp. $\inf_{Y \in \mathfrak{X}} D(Y)$).

Si X est un élément fixé dans $\mathfrak{X} - C$, l'événement $X \in C_n$ a une probabilité tendant vers zéro. En effet, $X \in C_n \Longrightarrow \inf_{Y \in \mathfrak{X}} D_n(Y) - D_n(X) = 0$, soit,

$$\inf_{Y \in \mathfrak{X}} D_n(Y) - \inf_{Y \in \mathfrak{X}} D(Y) + D(X) - D_n(X) = D(X) - \inf_{Y \in \mathfrak{X}} D(Y).$$

Or, X n'appartenant pas à C, le second membre est un nombre strictement positif soit 2ϵ. D'où

$$\Pr\{\inf_{Y \in \mathfrak{X}} D_n(Y) - D_n(X) = 0\} \leqslant \Pr\{\inf_{Y} D_n(Y) - \inf_{Y} D(Y) \geqslant \epsilon\} + \Pr\{D(X) - D_n(X) \geqslant \epsilon\}.$$

Or chacune des probabilités du second membre tend vers zéro ; et $\Pr\{X \in C_n\}$ est inférieur au premier membre.

L'événement complémentaire de $C_n \subset C$ peut être exprimé comme suit : $(\exists X)$: $X \notin C$ et $X \in C_n$; sa probabilité est majorée par : $\sum_{X \in \mathfrak{X} - C} \Pr\{X \in C_n\}$ somme finie tendant vers zéro.

B – DEUXIEME PROBLEME

1 — Introduction

Ici nous supposons fixée la population E d'objets. Chaque caractère a_h, $(E \xrightarrow{a_h} A_h)$, définit comme on l'a vu une partition S_h ; et les partitions centrales sont celles qui minimisent la fonction

$$D_p(X) = \frac{1}{p} \sum_{h=1}^{p} d(X, S_h) \qquad (1)$$

où X appartient à $\mathfrak{P}(E)$, ensemble des partitions sur E, et où $d(X, S_h)$ est le cardinal de la différence symétrique des graphes, dans $E \times E$, des relations d'équivalence définies par X et S_h. La suite des caractères $(a_1, a_2, \ldots, a_p)$ sera considérée comme un échantillon d'un ensemble plus vaste de caractères. D'une manière précise, nous supposerons que la famille de partitions, $\{S_h : h = 1, 2, \ldots, p\}$, est un échantillon, de p éléments aléatoires indépendants, pris dans $\mathfrak{P}(E)$ muni de l'algèbre de ses parties et d'une mesure de probabilité L.

$\{S_h : h = 1, 2, \ldots, p\}$ définit une mesure de probabilité empirique Lp, où $Lp(X)$ est la proportion de partitions S_h identiques à X. Relativement à la mesure Lp, l'expression (1) devient :

$$D_p(X) = \sum_{Y \in \mathfrak{X}} d(X, Y)\, Lp(Y) \qquad (2)$$

Le centre C_p, ensemble des partitions centrales sur E, relativement à la famille $\{S_h : h = 1, 2, \ldots, p\}$, est défini de la façon suivante :

$$C_p = \{X \in \mathfrak{P} : D_p(X) \leqslant D_p(Y) \text{ pour tout } Y \text{ de } \mathfrak{P}\}$$

D'une manière analogue, en posant,

$$D(X) = \sum_{Y \in \mathfrak{X}} d(X, Y)\, L(Y) \qquad (3)$$

on définira le centre C, ensemble des partitons de E où D atteint son minimum.

$$C = \{X \in \mathfrak{X} : D(X) \leqslant D(Y) \text{ pour tout } Y \text{ de } \mathfrak{X}\}$$

Pour X fixé, $D_p(X)$ est une variable aléatoire définie sur $\mathfrak{X}^p$ muni de l'algèbre de ses parties et de la mesure de probabilité induite par L. Au vecteur aléatoire $(D_p(X) : X \in \mathfrak{X})$ est attaché le centre C_p qui devient une partie aléatoire de $\mathfrak{X}$. Ici encore c'est au comportement asymptotique de

$$C_p, (p \longrightarrow \infty)$$

que nous allons nous intéresser ; de ce fait, certains des événements considérés se situerons dans $\prod_{i=1}^{\infty} \mathfrak{X}_i$ muni de la σ – algèbre $\bigotimes_{i=1}^{\infty} P(\mathfrak{X}_i)$ où $\mathfrak{X}_i = \mathfrak{X}$ pour tout i et où $P(\mathfrak{X}_i)$ est l'ensemble des parties de $\mathfrak{X}_i$.

Ce problème peut être inscrit dans le cadre plus général où la donnée est un espace de probabilité $(\underline{\Omega}, \mathfrak{A}, L)$, muni d'une distance $d(x, y)$, telle que les applications $y \longrightarrow d(x, y)$ de $\underline{\Omega}$ dans R soient mesurables pour tout x de $\underline{\Omega}$ (c'est-à-dire, telle que toutes les boules soient mesurables). Pour m strictement positif, donné, on peut définir (Fréchet) le moment d'ordre m par rapport au point x

$$D(x) = \int d^m(x, y)\, L(dy) \qquad (4)$$

qui est supposé partout fini ; sinon, l'inégalité triangulaire permet de voir que $D(x)$ est partout infinie.

Le centre C d'ordre m serait

$$C = \{x \in \underline{\Omega} : D(x) \leqslant D(y) \text{ pour tout } y \text{ de } \underline{\Omega}\}$$

Si $\underline{\Omega}$ est fini, comme c'est le cas pour notre problème, ou compact pour la métrique d, C est non vide. Si $\underline{\Omega}$ est un espace vectoriel sur R et si la métrique est euclidienne, l'unique valeur centrale d'ordre 2 est donnée par l'espérance mathématique

$$E(x) = \int x\, L(dx) \qquad (5)$$

et ceci en vertu de l'identité de Kœnig :

$$\int d^2(x, y)\, L(dy) = \int d^2(E(x), y)\, L(dy) + d^2(x, E(x)). \qquad (6)$$

Soit

$$D_p(x) = \frac{1}{p} \sum_{h=1}^{p} d^m(x, x_h)$$

où $(x_1, x_2, \ldots, x_p)$ est un échantillon de p éléments aléatoires indépendants de $\underline{\Omega}$. A $D_p(x)$ est attaché le centre C_p où D_p atteint sa borne inférieure. On peut, ici, formuler d'une manière plus générale le problème de convergence, posé ci-dessus, que nous allons reprendre dans le cadre précédent.

2 – Proposition 1

On a presque sûrement : lim sup $C_p \subset C$, c'est-à-dire :

$$\Pr.\{\limsup C_p \subset C\} = 1 \; ; \qquad (1)$$

d'autre part $$\lim_{p \to \infty} \Pr\{C_p \subset C\} = 1 \qquad (2)$$

La démonstration de cette proposition est strictement analogue à celle du paragraphe A ; en effet, on a bien, pour X fixé dans $\mathfrak{X}$, la convergence presque sûre de la suite de variables aléatoires, $D_p(X)$, vers $D(X)$.

Nous allons à présent pouvoir établir un résultat qui complète celui, (1), de la proposition ci-dessus ; ce résultat tient compte de plus près de la nature de la statistique

$$D_p(X) = \frac{1}{p} \sum_{i=1}^{p} d(X, X_i)$$

et en particulier de l'indépendance des éléments aléatoires Xi, $i = 1, 2, \ldots, p$.

3 – Proposition 2

On a presque sûrement $C \subset \limsup C_p$ soit $\Pr.\{C \subset \limsup C_p\} = 1$.

Démonstration

L'événement $\{C \subset \limsup C_p\}$ peut être exprimé sous la forme $\bigcap_{X \in C} \{X \in \limsup C_p\}$.

Dans ces conditions, il suffit de montrer que, pour X appartenant à C, $\Pr\{X \in \limsup C_p\} = 1$.

Or $\Pr\left\{X \in \bigcap_k \bigcup_{p>k} C_n\right\} = \lim_k \searrow \Pr\left\{X \in \bigcup_{p>k} C_p\right\}$, la limite étant décroissante, il y a lieu de montrer que pour tout entier k $\Pr\left\{X \in \bigcap_{p>k} C_p\right\} = 1$. Il revient au même de prouver que $\Pr\left\{X \in \bigcap_{p>k} C_p^c\right\} = 0, C_p^c$ étant le complémentaire de C_p dans $\mathfrak{R}$, c'est-à-dire :

$$\Pr\left\{\bigcap_{p>k} \bigcup_{Y \in \mathfrak{R}} \{D_p(Y) - D_p(X) < 0\}\right\} = 0. \quad (1)$$

Pour tout Y de $\mathfrak{R}$, posons

$$S_p(Y) = p(D_p(Y) - D_p(X)) = \sum_{i=1}^{p} (d(Y, X_i) - d(X, X_i)) \quad (2)$$

c'est une somme de variables,

$$\delta_i(Y) = d(Y, X_i) - d(X, X_i),$$

mutuellement indépendantes et de même loi, bornées par

$$B = \max_Y \max_Z (d(Y, Z) - d(X, Z)), \quad (3)$$

Y et Z de $\mathfrak{R}$, B est une borne *finie* puisque $\mathfrak{R}$ est fini.

L'espérance mathématique de $\delta_i(Y)$ est :

$$\delta(Y) = E(\delta_i(Y)) = D(Y) - D(X) \geqslant 0,$$

(X appartenant à C) et sa variance, qui est finie, sera notée V.

Considérons à présent la suite infinie d'entiers, (p_j), telle que $p_1 = h + 1$ et $p_{j+1} = p_j + l^2 p_j^2$, où l est un paramètre entier que nous nous fixons, et introduisons pour chaque Y de $\mathfrak{R}$,

$$Z_j(Y) = S_{p_{j+1}}(Y) - S_{p_j}(Y). \quad (4)$$

$Z_j(Y)$ est une somme de $q_j = p_{j+1} - p_j$ variables indépendantes ; d'autre part, les $Z_j(Y)$ sont mutuellement indépendants, car les sommes qui les définissent portent, respectivement, sur des ensembles d'indices disjoints.

Reprenons la formule (1) à démontrer. On a :

$$\bigcap_{p>k} \bigcup_{Y \in \mathfrak{R}} \{D_p(Y) - D_p(X) < 0\} \subset \subset \bigcap_{j=1}^{\infty} \bigcup_{Y \in \mathfrak{R}'} \{Z_j(Y) \leqslant Bp_j\} \quad (5)$$

où B est défini par (3) et où $\mathfrak{R}' = \mathfrak{R} - \{X\}$.

En effet ;

$$S_{p_{j+1}}(Y) < 0 \Longleftrightarrow Z_j(Y) < -S_{p_j}(Y)$$

et d'autre part $-S_{p_j}(Y) \leqslant Bp_j$. Donc la probabiblité du premier membre de (5) est inférieure ou égale à

$$\Pr\left\{\bigcap_{j=1}^{\infty}\left[\bigcup_{Y \in \mathfrak{R}'} \{Z_j(Y) \leqslant Bp_j\}\right]\right\} = \prod_{j=1}^{\infty} \Pr.\left\{\bigcup_{Y \in \mathfrak{R}'} \{Z_j(Y) \leqslant Bp_j\}\right\}$$

en vertu de l'indépendance des $Z_j(Y)$.

Un tel produit infini est nul si

$$\lim_{j \to \infty} \Pr\left\{\bigcup_{Y \in \mathfrak{R}'} \{Z_j(Y) \leqslant Bp_j\}\right\} < 1. \quad (6)$$

Or, si j tend vers l'infini, $q_j = l^2 . p_j^2$ tend encore vers l'infini, la loi limite de

$$(Z_j(Y) - q_j \delta(Y))/\sqrt{q_j} = T_j(Y) \quad (7)$$

est, en vertu du théorème central limite, normale, de moyenne zéro et de variance $V = \text{Var.}(\delta_i(Y)$ en tenant compte de $q_j = l^2 . p_j^2$, on a :

$$\{Z_j(Y) \leqslant Bp_j\} = \left\{T_j(Y) \leqslant \frac{B}{l} - \delta(Y)\, l . p_j\right\}. \quad (8)$$

Il en résulte que

a) $\lim_{j \to \infty} \Pr\{Z_j(Y) \leqslant Bp_j\} = 0$ si Y n'appartient pas à C, soit $\delta(Y) > 0$.

b) $\lim_{j \to \infty} \Pr\{Z_j(Y) \leqslant Bp_j\} = \lim_{j \to \infty} \Pr.\left\{T_j(Y) \leqslant \frac{B}{l}\right\}$

si Y appartient à C, soit $\delta(Y) = 0$. Le dernier second membre est strictement inférieur à 1.

Reprenons le premier membre de (6) ; en vertu de (a), $\Pr\left\{ \bigcup_{Y \in \mathfrak{C}'} \{Z_j(Y) \leqslant Bp_j\} \right\}$ a une limite inférieure ou égale à celle de

$$\Pr\left\{ \bigcup_{Y \in C'} \left\{ T_j(Y) \leqslant \frac{B}{l} \right\} \right\} \qquad (9)$$

où $C' = C - \{X\}$.

Or le vecteur $\{T_l(Y),\ Y \in C'\}$ suit asymptotiquement une loi normale de moyenne, de vecteur nul ; et dont la matrice de variances et covariances est finie. La limite, pour j tendant vers l'infini, de la probabilité (9), est donc, *en général*, strictement plus petite que 1.

La probabilité (9) ne peut tendre vers 1, pour tout l, que si la loi asymptotique de $T_j(Y)$ est concentrée au point 0 pour un Y de C'. Il est nécessaire pour cela d'avoir presque sûrement

$$d(X, V) = d(Y, V)$$

pour V aléatoire dans $\mathfrak{C}$. En d'autres termes

$$\Pr\{V : d(X, V) \neq d(Y, V)\} = 0\ , \qquad (10)$$

Y étant différent de X, $(Y \in C - X)$, pour $V = X$ on a $0 = d(X, X) \neq d(Y, X)$ donc si la masse affectée à X par L est non nulle, (10) est impossible.

III — *H-CLASSIFICABILITE*

1 — Introduction

De ce nom, nous désignons l'aptitude d'une population E d'objets à être organisée en une hiérarchie de classifications respectant d'une manière satisfaisante les ressemblances entre objets, (nous dirons plus brièvement à être h-classifiée).

L'ensemble A des attributs, [cf. Ch.1 § I-1.1], est ici fixé une fois pour toutes et E est considéré comme un échantillon d'une population E' plus vaste que E. Nous nous placerons ici dans le cadre où la donnée de base est une préordonnance sur E, ω.

Le premier problème que devrait se poser le taxinomiste est celui de juger si, au vu des données (ici de ω), il n'est pas artificiel d'imposer sur E une structure de chaîne de partitions censée respecter les ressemblances entre objets. Expliquons-nous sur ce point à l'aide d'une illustration :

"Si E est un ensemble de points pris, "au hasard", selon une loi uniforme dans le plan euclidien ; le problème de chercher une h-classification sur E, muni de la distance euclidienne ou de l'ordonnance associée, n'a aucun intérêt. La seule hiérarchie de classifications admissible dans ce cas est la plus grossière : celle à deux niveaux, dont le premier définit la partition la plus fine sur E et le second, la moins fine".

Le problème posé est celui de trouver un test assez puissant permettant d'éprouver l'hypothèse de "l'indépendance des objets" relativement à leur description, soit de la "non existence d'une h-classification naturelle". Nous préciserons plus loin cette hypothèse et le test. Mais faisons remarquer que la nature statistique de problème posé est très classique. En effet, par exemple, pour un ensemble de points observés dans le plan euclidien, il existe bien un test statistique qui éprouve l'hypothèse de l'indépendance des points au profit de celle de la régression linéaire. Un tel test est à effectuer avant l'ajustement de l'ensemble des points par une droite.

A notre connaissance, ce problème n'a guère préoccupé les taxinomistes, . . . peut-être parce que le principal problème est d'estimation (trouver la meilleure h-classification !).

Notons enfin que si (E, ω) a un degré d'aptitude suffisant à être h-classifié, on peut s'attendre à ce qu'un algorithme, donné sous-optimal pour un critère, fournisse la préordonnance ultramétrique optimale, vis-à-vis de l'un quelconque des critères. Cette considération intuitive est renforcée par le fait qu'il arrive souvent que différentes méthodes fournissent des hiérarchies de classifications très voisines.

Nous commencerons par proposer une statistique pour mesurer la h-classificabilité de (E, ω). L'étude de la distribution de notre statistique dans l'hypothèse, que nous préciserons, de la "non existence

d'une h-classification naturelle" nous permettra de proposer un test qui éprouve cette hypothèse.

2 – Ecartement entre la structure de ω et celle d'une préordonnance ultramétrique

Une condition nécessaire et suffisante pour qu'une distance d, définie sur E, soit ultramétrique est que tout triangle soit isocèle, la base étant le plus petit des côtés. Cette proposition devient, relativement à la préordonnance ω : ω est ultramétrique, si et seulement si, pour tout triplet $\{x, y, z\}$, (partie à trois éléments de E), pour lequel

$$\{x, y\} \leqslant \{y, z\} \leqslant \{x, z\},$$

on a encore $\{x, z\} \leqslant \{y, z\}$, c'est-à-dire, $\{x, z\}$ et $\{y, z\}$ dans une même classe du préordre ω.

Soit dans ces conditions J l'ensemble des parties à trois éléments de E ; considérons l'application τ de J dans l'ensemble des intervalles ouverts de F, pour ω ; qui, à chaque triplet x, y, z associe l'intervalle ouvert

$$]M(x, y, z), S(x, y, z)[\quad \text{où} \quad M(x, y, z)$$

désigne la paire médiane et $S(x, y, z)$ la paire de rang le plus élevé parmi les trois paires

$$\{x, y\}, \{y, z\} \quad \text{et} \quad \{z, x\}.$$

Une telle application peut être matérialisée par un tableau T dont les colonnes portent les éléments de F dans l'ordre où ils se présentent pour ω, (ainsi chaque colonne porte le nom d'une paire et les différentes colonnes correspondant à une même classe du préordre ω sont regroupées et séparées de celles correspondant à la classe suivante). Chaque ligne du tableau T représente un élément de J. Le tableau est à $n(n-1)/2$ colonnes et

$$n(n-1)(n-2)/6$$

lignes, ($|E| = n$). Chaque intervalle du type

$$]M(x, y, z), S(x, y, z)[$$

est porté par la ligne du tableau x, y, z .

Exemple : soit $E = \{a, b, c, d, e\}$ et ω la préordonnance sur E

$$\{a, d\} = \{a, c\} < \{a, e\} < \{c, e\} < \{b, d\} =$$

$$= \{c, d\} < \{b, c\} < \{d, e\} < \{a, b\} < \{b, e\}$$

Pour une même ligne correspondant par exemple à $\{x, y, z\}$, les trois points indiquent les trois paires $\{x, y\}$, $y, z\}$ et $\{x, z\}$; la suite des croix l'intervalle $]M(x, y, z), S(x, y, z)[$.

Une mesure de l'"écartement" entre la structure de ω et celle d'une préordonnance ultramétrique va être définie à partir d'une mesure sur F, dépendant de ω.

Tableau T

J \ F	(a, d)	(a, c)	(a, e)	(c, e)	(b, d)	(c, d)	(b, c)	(d, e)	(a, b)	(b, e)
(a, b, c)		•					•	×	•	
(a, b, d)	•				•		×	×	•	
(a, b, e)			•						•	•
(a, c, d)	•	•	×	×		•				
(a, c, e)		•	•	•						
(a, d, e)	•		•	×	×	×	×	•		
(b, c, d)					•	•	•			
(b, c, e)				•			•	×	×	•
(b, d, e)					•			•	×	•
(c, d, e)				•		•	×	•		

2.1 — *Statistiques mesurant l'écartement*

Nous venons de dire qu'un triplet $\{x, y, z\}$ pour lequel on a $\{x, y\} \leqslant \{y, z\} \leqslant \{x, z\}$, pour ω, est tel que l'intervalle $]\{y, z\}, \{x, z\}[$ est vide si ω est ultramétrique. Vis à vis d'un tel triplet, la préordonnance ω est d'autant "moins ultramétrique" que le cardinal de $]\{y, z\}, \{x, z\}[$ défini sur ω, est grand. Pour tenir compte de l'ensemble J de tous les triplets, on peut adopter comme mesure de l'"écartement" entre la structure de ω et celle d'une préordonnance ultramétrique :

$$HC(\omega) = \sum_J |]M(x, y, z), S(x, y, z)[|$$

$$= \sum_J \Delta(x, y, z)$$

où $\Delta(x, y, z)$ est le cardinal :

$$|]M(x, y, z), S(x, y, z)[|$$

($|J| = n(n-1)(n-2)/6$ si $|E| = n$).

Référons-nous à la représentation de l'application τ par le tableau T. $HC(\omega)$ est le nombre de croix dans le tableau. $HC(\omega)$ a été définie en lisant T ligne par ligne. Si nous regardons le tableau verticalement (colonne par colonne), $HC(\omega)$ apparaît comme une mesure sur F, affectant à chaque élément p de F, $|J_p|$ où J_p est le sous-ensemble de J dont chaque élément $\{x, y, z\}$ est tel que

$$M(x, y, z) < p < S(x, y, z),$$

pour ω. En d'autres termes $|J_p|$ est le nombre de triplets pour lesquels la paire médiane et la paire supérieure enserrent strictement p ou encore le nombre de fois où la paire p intervient pour séparer strictement une paire $M(x, y, z)$ d'une paire $S(x, y, z)$.

Pour rendre la statistique $HC(\omega)$ indépendante du cardinal de E nous la réduisons en remplaçant toute mesure absolue (cardinal) par une mesure relative (proportion), nous obtenons ainsi

$$H(\omega) = \frac{1}{|J|} \sum_J \frac{|]M(x, y, z)\, S(x, y, z)[|}{|F|} =$$

$$= HC(\omega)/|F| \times |J|$$

La lecture du tableau T colonne par colonne permet d'écrire, avec les notations de ci-dessus,

$$H(\omega) = \frac{1}{|F|} \sum_{p \in F} \frac{|J_p|}{|J|}$$

$H(\omega)$ apparaît comme la moyenne de la distribution sur F qui à chaque élément p de F attache la pondération $|J_p|/|J|$. D'où l'idée de caractériser la classificabilité de E par la distribution des rationnels positifs $m_p = |J_p|/|J|$, $p \in F$, exactement par la suite décroissante de ces nombres ; une telle suite sera notée $D(\omega)$.

Dans l'exemple ci-dessus :

$$H(\omega) = 13/10 \times 10 = 0.13$$

et $D(\omega) = (0.3, 0.3, 0.2, 0.2, 0.1, 0.1, 0.1, 0, 0, \ldots)$.

Si ω est ultramétrique, $H(\omega)$ est nul et le vecteur $D(\omega)$ a toutes ses composantes nulles. $H(\omega)$ permet de préordonner totalement l'ensemble $\mathfrak{E}$ des couples (E, ω), où E est un ensemble fini et ω une préordonnance sur E. $D(\omega)$ fournit sur l'ensemble $\mathfrak{E}$ un préordre seulement partiel, néanmoins plus sûr que le préordre total précédent ; défini de la façon suivante : $(E, \omega) \leqslant (E', \omega')$ (*i.e.* (E, ω) plus classifiable que (E', ω') dans $\mathfrak{E}$) si et seulement si $D(\omega) \leqslant D(\omega')$ (i.e. $d_i \leqslant d'_i$ pour tout i, où d_i (resp. d'_i) est le $i^{\text{ème}}$ élément de la suite $D(\omega)$ (resp. $D(\omega')$)).

Il est manifeste que le préordre total défini à partir de $H(\omega)$ est compatible avec ce préordre partiel :

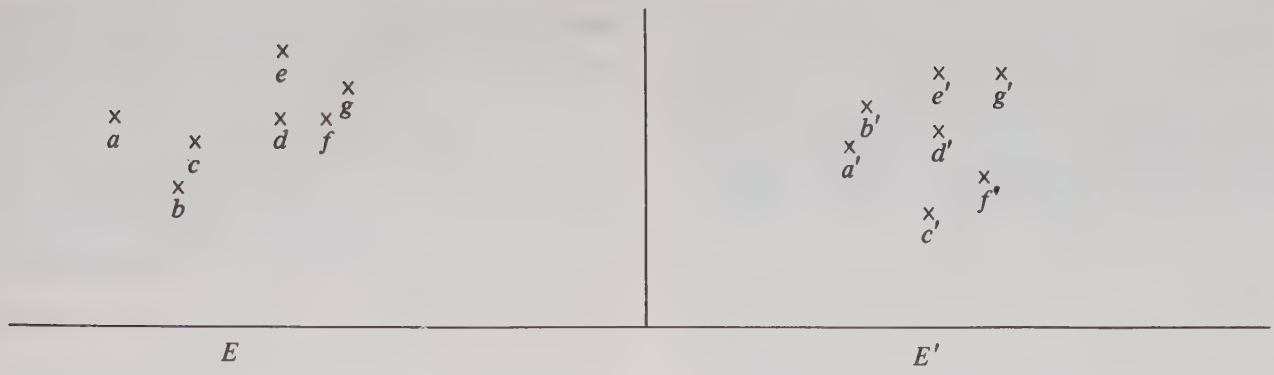
$$D(\omega) \leqslant D(\omega') \Longrightarrow H(\omega) \leqslant H(\omega').$$

2.1.1 — *Illustration géométrique*

Nous figurons ci-dessous deux ensembles E et E' de 7 points chacun dans le plan euclidien.

$$E = \{a, b, c, d, e, f, g\},\ E' = \{a', b', c', d', e', f', g'\}.$$

L'ensemble E (resp. E') est muni de la préordonnance ω (resp. ω') associée à la distance euclidienne définie sur E, (resp. sur E'). Pour un même ensemble, le nombre de paires de points distincts est 21 et le nombre de triplets 35.



La préordonnance ω

$(d, f) = (f, g) < (b, c) < (d, e) = (d, g) < (e, g) < (e, f) < (a, c) < (c, d) < (a, b) < (b, d) < (c, e) = (c, f) < (b, f) = (c, g) < (a, d) < (b, e) < (a, e) < (b, g) < (a, f) < (a, g)$.

La préordonnance ω'

$(a, b) < (e, g) < (d, e) < (c, f) = (d, f) < (b, d) < (a, d) < (b, e) = (c, d) < (d, g) < (a, c) < (f, g) < (e, f) = (a, e) < (b, c) < (a, f) < (b, f) = (b, g) < (c, e) < (a, g) = (c, g)$.

Il semble, visuellement, que (E, ω) soit "mieux classifiable" que (E', ω'). En effet on pourrait distinguer pour E la classification $(\{a, b, c\}, \{d, e, f, g\})$.

On a trouvé pour (E, ω) , $H(\omega) = 77/21 \times 35$

$$D(\omega) = \left(\frac{10}{35}, \frac{10}{35}, \frac{9}{35}, \frac{7}{35}, \frac{7}{35}, \frac{7}{35}, \frac{6}{35}, \frac{6}{35}, \frac{4}{35}, \frac{4}{35}, \frac{4}{35}, \frac{1}{35}, \frac{1}{35}, \frac{1}{35}, 0, 0, \cdots\right)$$

et pour (E', ω') , $H(\omega') = 128/21 \times 35$

$$D(\omega') = \left(\frac{14}{35}, \frac{13}{35}, \frac{13}{35}, \frac{12}{35}, \frac{11}{35}, \frac{11}{35}, \frac{9}{35}, \frac{9}{35}, \frac{9}{35}, \frac{7}{35}, \frac{5}{35}, \frac{5}{35}, \frac{5}{35}, \frac{3}{35}, \frac{1}{35}, \frac{1}{35}, 0, 0, \cdots\right)$$

2.2 – *Simulation statistique de* $H(\omega)$ et de $D(\omega)$

Si n est grand, $|J|$ peut être trop important. Nous simulerons $H(\omega)$ au moyen d'un échantillon Q de q tirages indépendants dans J muni d'une mesure uniforme, (équiprobabilité d'apparition de chacun des triplets à chaque tirage). La statistique obtenue sera désignée par :

$$H'(\omega) = \frac{1}{q} \sum_{\{x, y, z\} \in Q} \Delta(x, y, z)/f$$

où $f = |F|$.

Pour q assez grand, $(H'(\omega) - H(\omega)/\sqrt{q}$ est, en vertu du théorème central limite, une variable aléatoire normale de moyenne nulle et de variance V :

$$V = \text{Var}\,(\Delta(x, y, z)/f)$$

Pour simuler la distribution $D(\omega)$, soit un échantillon de k tirages indépendants dans F muni d'une mesure uniforme (équiprobabilité d'apparition de chacune des paires à chaque tirage). Si p est un élément tiré, la statistique m_p correspondante (qu'il y a lieu de calculer) est une valeur observée de la variable qu'on vise. La suite des k valeurs ainsi obtenues définira la distribution empirique simulant $D(\omega)$.

3 – Test de l'hypothèse N de la non existence d'une h-classification "naturelle".

Nous allons commencer par préciser une telle hypothèse. Rappelons que E nous est donné comme un échantillon dans $P(A)$, ensemble des parties de

l'ensemble des attributs. Sur $P(A)$ les objets d'une population "bien classifiable" apparaissent comme des "îlots sporadiques". Il semble par conséquent naturel d'exiger que, pour l'hypothèse N, E corresponde à un échantillon de n éléments aléatoires indépendants dans $P(A)$ muni d'une mesure adéquate. Notre première réaction a été de vouloir adopter sur $P(A)$ une mesure uniforme ; cependant si, par exemple, le nombre d'attributs possédés par un même objet de E était invariable, soit a, il est plus cohérent de concentrer toute la masse sur le niveau du simplexe $P(A)$ correspondant à a. Par conséquent, soit la partition de E selon le nombre d'attributs possédés par un même objet ;

$$E = E_1 + E_2 + \cdots + E_j + \cdots + E_k ,$$

où tout objet de E_j possède exactement a_j attributs. Si n_j est le cardinal de

$$E_j, (n_1 + n_2 + \cdots + n_k = n) ,$$

affectons au niveau du simplexe $P(A)$, formé des parties de A ayant a_j éléments, la masse n_j/n que nous répartissons d'une manière uniforme sur les différents éléments de ce niveau. L'hypothèse N ainsi exprimée semble assez bien refléter la non classificabilité. En effet, reportons nous au tableau d'incidence des données Objets x Attributs, $(E \times A)$, à n lignes et p colonnes, et considérons l'ensemble $\mathfrak{T}$ des tableaux $n \times p$ formés de 0 et de 1 pour lesquels le nombre de 1 de la ligne i est celui de la ligne correspondante du tableau $E \times A$; l'hypothèse N consiste à considérer notre tableau des données comme issu de $\mathfrak{T}$ muni du modèle uniforme.

On peut effectuer le test sur la base de la statistique $H(\omega)$ ou de la distribution $D(\omega)$. Considérons pour cela un échantillon E' de n éléments aléatoires indépendants, (tirage avec remise), dans $P(A)$ muni de la mesure de probabilité ci-dessus définie ; et établissons sur E' la préordonnance ω' de la même façon que l'a été ω sur E (avec le même indicateur de ressemblance). Dans la mesure où on connaît la distribution de $H(\omega)$ sur l'ensemble des échantillons de type E', on pourra effectuer le test au vu de la valeur de $H(\omega)$. Sinon, E' simulera l'hypothèse N et le test est défini à partir de $H'(\omega)$, [cf. § 2.2 ci-dessus]. Soit alors J' l'ensemble des parties à trois éléments de E', Q' un échantillon de q éléments aléatoires indépendants de J', muni d'une mesure uniforme ; enfin $G'(\omega)$ la statistique correspondant à $H'(\omega)$ calculée à partir de Q'. En vertu du théorème central limite, dans l'hypothèse N, $(H'(\omega) - G'(\omega))/\sqrt{q}$ suit approximativement, pour q grand, une loi normale de moyenne nulle et de variance $2V$ où

$$V = \mathrm{Var}\, \Delta(x, y, z)/f$$

qu'on estimera à partir de l'échantillon Q'.

Il semble plus satisfaisant d'effectuer le test en comparant les deux distributions $D(\omega)$ et $D(\omega')$ où $D(\omega')$ est associé à l'échantillon E' muni de ω', défini ci-dessus. Dans la mesure où $D(\omega)$ est à gauche de $D(\omega')$, $(D(\omega) < D(\omega'))$, on conclura à la classificabilité de la population étudiée.

Si n est grand, c'est sur des simulations de $D(\omega)$ et $D(\omega')$ qu'on basera le test. Il reste que les calculs sont assez importants et dépendent d'une simulation par échantillonnage ; d'où l'intérêt de connaître dans l'hypothèse N la distribution théorique des valeurs m_p , $p \in F$. Désignons par M la variable aléatoire associée.

3.1 — *Distribution de M dans un cas particulier important*

Ce cas particulier est celui où toute la masse de la mesure de probabilité, définie ci-dessus, est concentrée sur un même niveau du simplexe $P(A)$; il correspond à la situation où le nombre d'attributs possédés par un même objet est invariable dans la population étudiée. Ce cas est très important, car la préordonnance associée est un invariant quel que soit l'indicateur de ressemblance (fonction de s, u et v) choisi pour la définir. Le spécialiste peut se ramener à ce cas dans la mesure où il peut se contenter d'une famille de hiérarchies de classifications, chacune établie pour la sous-population d'objets ayant un nombre d'attributs fixé ; cela est d'autant plus aisé que le nombre d'attributs possédés par un même objet aura une faible variance dans la population étudiée.

$A = \{\alpha_1, \alpha_2, \ldots, \alpha_T\}$: au lieu de travailler dans $P(A)$ il nous est équivalent et plus commode de travailler dans $\{0, 1\}^T$; en fait, seul un hyperplan K de $\{0, 1\}^T$ nous intéressera, celui pour lequel $\sum_{i=1}^{T} x_i = k$ ($x_i = 0$ ou 1). On suppose une mesure de probabilité sur K équirépartie. Tous les indices de similarité étant équivalents sur K, considérons la préordonnance associée à l'indice $s = \sum_{i=1}^{T} x_i y_i$.

Les différentes classes de la préordonnance sont définies à partir des différentes valeurs possibles de l'indice $s = \sum_i x_i y_i$; l'ensemble de ces valeurs étant $\{0, 1, 2, \ldots, m, \ldots, k\}$. Indexons chaque classe du préordre total par la valeur de l'indice correspondant. Une des valeurs de la variable aléatoire qui nous intéresse est définie à partir d'une classe donnée comme étant la proportion de triplets dont la paire médiane et la paire supérieure enserrent strictement la classe en question ; la pondération ou probabilité attachée à cette valeur de la variable est la proportion des paires qui appartiennent à la classe. $M(m)$ désignera la valeur de la variable associée à la classe m, $P(m)$ la pondération affectée à cette valeur de la variable.

Dans ce contexte

$$H(\omega) = \sum_{m=0}^{k} M(m) \cdot P(m)$$

et $D(\omega)$ la suite *décroissante* des valeurs $M(m)$, *pondérées* respectivement par $P(m)$;

$$m = 0, 1, 2, \ldots, k .$$

Dans l'hypothèse statistique précisée ci-dessus, occupons-nous de la recherche des nombres $M(m)$ et de leurs pondérations respectives.

$$M(m) = \sum_{HK} 3 \Pr\left\{ \sum_i x_i y_i = p \, , \sum_i y_i z_i = q \, , \sum_i x_i z_i = r \right\} \qquad (1)$$

où HK est l'ensemble

$$HK = \{(p, q, r) \mid p < m, q < m, r > m$$

et

$$p + q \leqslant r + k, q + r \leqslant p + k, r + p \leqslant q + k\}$$

et où x, y, z sont trois points indépendants pris dans K selon une loi uniforme ; X, Y, Z les sous-ensembles d'attributs correspondants. Dans l'expression de HK, les inégalités de type

$$p + q \leqslant r + k$$

sont impliquées par celles de type

$$\sum_i (x_i - z_i)^2 \leqslant \sum_i (x_i - y_i)^2 + \sum_i (y_i - z_i)^2$$

avec

$$\sum_i x_i = \sum_i y_i = \sum_i z_i = k$$

$M(m)$ est évidemment nul pour $m = 0$ et $m = k$.

$P(m)$ est la proportion des paires $\{x, y\}$ de K pour lesquelles $\sum_i x_i y_i = m$; soit

$$P(m) = \Pr\{\Sigma x_i y_i = m\} = \binom{k}{m}\binom{T-k}{k-m} \Big/ \binom{T}{k} \qquad (1')$$

Le terme général de la somme (1) peut se mettre sous la forme

$$\Pr\left\{ \sum_i x_i y_i = p \, , \sum_i y_i z_i = q \Big/ \sum x_i z_i = r \right\} \Pr\left\{ \sum_i x_i z_i = r \right\} \qquad (2)$$

or

$$\Pr\left\{ \sum_i x_i z_i = r \right\} = \binom{k}{r}\binom{T-k}{k-r} \Big/ \binom{T}{k} \qquad (3)$$

et l'autre facteur du produit (2) vaut

$$\sum_{l=0}^{r} \binom{k-r}{p-l}\binom{k-r}{q-l}\binom{r}{l}\binom{T-(2k-r)}{k-(p+q-l)} \Big/ \binom{T}{k} \qquad (4)$$

Un terme courant de cette somme représente la probabilité que les k attributs de Y se répartissent de la manière suivante : l attributs sont dans $X \cap Z$, $(p - l)$ dans $X - Z$, $(q - l)$ dans $Z - X$ et le restant soit $k - (p + q - l)$ dans $A - X \cup Z$.

D'où l'expression de $M(m)$

$$M(m) = \sum_{HK} 3 \frac{\sum_{l=0}^{r} \binom{k-r}{p-l}\binom{k-r}{q-l}\binom{r}{l}\binom{T-(2k-r)}{k-(p+q-l)}}{\binom{T}{k}}$$

$$\times \binom{k}{r}\binom{T-k}{k-r}\binom{T}{k} \qquad (5)$$

Compte tenu de la définition de HK, $M(m)$ est obtenu à partir d'une sommation d'ordre 4 ; nous avons programmé le calcul de l'expression ci-dessus, le temps de calcul est essentiellement une fonction de k ; ce temps reste raisonnable pourvu que k ne dépasse pas 75 ; au-delà de $k = 100$, il devient par trop important. D'où l'intérêt de trouver une formule approchée permettant d'éviter, par exemple, la somme sur l ; cette dernière s'effectuant analytiquement.

Considérons par conséquent dans l'expression (5), la somme entre parenthèses, elle peut se mettre sous la forme :

$$\frac{\binom{k}{p}\binom{T-k}{k-p}}{\binom{T}{k}} \times \sum_{l=0}^{r} \frac{\binom{k-r}{p-l}\binom{r}{l}}{\binom{k}{p}} \times \frac{\binom{k-r}{q-l}\binom{T-(2k-r)}{k-(p+q-l)}}{\binom{T-k}{k-p}} \qquad (6)$$

Un terme de cette somme est le produit de deux probabilités de type hypergéométrique. Bénéficiant de l'approximation de ce type de probabilité par celui de Poisson ; on a, en posant

$$\mu = pr/k \quad \text{et} \quad \nu = (k-p)(k-r)/(T-k)$$

$$\binom{k-r}{p-l}\binom{r}{l} \Big/ \binom{k}{p} \simeq \frac{\mu^l}{l!} e^{-\mu}$$

et

$$\binom{k-r}{q-l}\binom{T-(2k-r)}{k-(p+q-l)} \Big/ \binom{T-k}{k-p} \simeq \frac{\nu^{q-l}}{(q-l)!} e^{-\nu}$$

La somme de l'expression (6) devient

$$\sum_{l=0}^{r} \frac{\mu^l}{l!} e^{-\mu} \frac{\nu^{q-l}}{(q-l)!} e^{-\nu} = \frac{e^{-(\mu+\nu)}}{q!} \sum_{l=0}^{r} \binom{q}{l} \mu^l \nu^{q-l}$$

Compte tenu du fait qu'une telle quantité ne sera à prendre en considération que pour $r > q$ (cf. l'expression de HK) , elle vaut

$$\frac{(\mu+\nu)^q}{q!} e^{-(\mu+\nu)}$$

Une valeur approchée de $M(m)$ est donc

$$M'(m) = \sum_{HK} 3 \frac{\binom{k}{p}\binom{T-k}{k-p}}{\binom{T}{k}} \times \frac{\binom{k}{r}\binom{T-k}{k-r}}{\binom{T}{k}} \times \frac{(\mu+\nu)^q}{q!} e^{-(\mu+\nu)} \qquad (7)$$

Une telle approximation est tout à fait valable comme en témoignent les tableaux suivants qui résultent d'un programme de calculs. Nous avons constaté par l'expérience (et on doit pouvoir le montrer aisément) que la distribution de $M(m)$ (resp. $M'(m)$), m variant par valeurs croissantes de 1 à $(k-1)$, était d'abord croissante sur une section commençante des valeurs de m puis décroissante sur la section finissante complémentaire, (distribution unimodale) ; les valeurs de $M(m)$ (resp. $M'(m)$) devenant négligeables sur une assez large section finissante des valeurs de m. Il nous a par conséquent semblé judicieux d'arrêter les calculs de $M(m)$ (resp. $M'(m)$) dès que les contributions de $M(m)\,P(m)$ à H et de $M'(m)\,P(m)$ à H' devenaient négligeables, (de manière précise dès que

$$M(m)\,P(m) < 10^{-3} \sum_{1}^{(m-1)} M(n)\,.\,P(n)$$

et

$$M'(m)\,P(m) < 10^{-3} \sum_{1}^{(m-1)} M'(n)\,.\,P(n))$$

Pour illustrer l'intérêt de la formule approchée, signalons que, pour une programmation FORTRAN sur IBM 360-75, nous avons observé les temps suivants

Distribution	exacte	approchée
$T = 250\,,\ k = 50$	40 s	14 s
$T = 246\,,\ k = 75$	250 s	51 s

$T = 100$, $k = 20$

m	$M(m)$	$M'(m)$	$P(m)$
1	0.000	0.000	0.043
2	0.007	0.013	0.126
3	0.064	0.084	0.216
4	0.186	0.201	0.244
5	0.222	0.217	0.192
6	0.134	0.127	0.109
7	0.048	0.045	0.046
8	0.011	0.011	0.014
9	0.002	0.002	0.003
S	0.674	0.699	0.994

$H = 0.120$ $H' = 0.127$

$\sum_m |M'(m) - M(m)| = 0{,}056$

$|H' - H| = 0.007$

$T = 150$, $k = 30$

m	$M(m)$	$M'(m)$	$P(m)$
1	0.000	0.000	0.005
2	0.000	0.000	0.024
3	0.003	0.005	0.067
4	0.024	0.036	0.130
5	0.099	0.122	0.185
6	0.212	0.226	0.201
7	0.257	0.252	0.170
8	0.195	0.184	0.115
9	0.100	0.095	0.062
10	0.038	0.036	0.028
11	0.011	0.011	0.010
S	0.939	0.968	0.996

$H = 0.138$ $H' = 0.144$

$\sum_m |M'(m) - M(m)| = 0.075$

$|H' - H| = 0.006$

$T = 200$, $k = 40$

m	$M(m)$	$M'(m)$	$P(m)$
1	0.000	0.000	0.001
2	0.000	0.000	0.004
3	0.000	0.000	0.014
4	0.001	0.002	0.038
5	0.008	0.015	0.078
6	0.042	0.060	0.126
7	0.125	0.148	0.164
8	0.229	0.242	0.174
9	0.277	0.273	0.154
10	0.234	0.223	0.114
11	0.146	0.138	0.071
12	0.071	0.067	0.038
13	0.028	0.027	0.017
14	0.009	0.009	0.007
S	1.170	1.203	0.997

$H = 0.149$ $H' = 0.155$

$\sum_m |M'(m) - M(m)| = 0.091$

$|H' - H| = 0.006$

ANNEXE I

PROGRAMMES

Nous reproduisons ici l'essentiel des programmes relatifs aux algorithmes de classification développés au chapitre 3.

Le langage utilisé est le Fortran.

Les spécifications particulières de forme des données (FORMAT), ou de contraintes physiques (DIMENSION) ainsi que les sorties à fin d'édition (titrage, disposition) ont été supprimées. Nous n'avons pas non plus fait figurer certains sous-programmes à usage purement technique (Gestion de tambour, Triclassement etc.).

Le programme suivant, intitulé RF.03, est relatif à l'algorithme des transferts ; il a été composé par Madame M. Renaud. Précisons qu'à partir d'une matrice de similarités et d'une classification initiale en P classes, ce programme transfère les objets d'une classe vers une autre en cherchant à maximiser une fonction linéaire VL des similarités.

Lorsqu'il n'est plus possible d'augmenter VL en P classes, le programme teste s'il est possible de l'améliorer en augmentant le nombre de classes ; si oui, P progresse de DP et ceci jusqu'à $P = MP$.

(P, DP et MP sont des paramètres à fixer par l'utilisateur).

Les résultats imprimés sont :

pour chaque élément son "attraction" avec chacune des classes de la classification initiale ;

la valeur maximum de VL obtenue et la classification qui lui correspond ;

pour chaque objet son "attraction" avec chacune des classes de la classification finale.

Pour ce programme qui comporte plusieurs options nous n'avons fait figurer que celle qui traite les données en un seul passage.

RF. 03

```
-D  RUN  RENAUD,TRANS
- ASG A=FC012
-I  FOR TRANS
C      NI=NOMBRE D'OBJETS
C      S=MATRICE DE SIMILARITE,REPARTIE EN-NN-BLOCS DE-II-LIGNES
C      NNXII=NI
C      IA=CLASSEMENT INITIAL
C      P=NOMBRE REEL DE CLASSE
C      DP=PAS DE PROGRESSION DE P
C      MP=NOMBRE MAXIMUM DE CLASSE
C      IH=TEMPS IMPARTI AU PROGRAMME(EN MINUTE)
C      IHD=DUREE DU PROGRAMME
C      IEEUR=TEST DE TEMPS DEPASSE
C      ITRANS=NOMBRE DE TRANSFERTS
```

```
C      KO=NOMBRE MAXIMUM DE TRANSFERTS
C      C=MATRICE DES ATTRACTIONS
C      VL=FORME LINEAIRE A MAXIMISER
C      GO=-NI(NI-1)/2 =VL INITIAL
C      D=MAXIMUM DE LA DIFFERENCE DES ATTRACTIONS DE L OBJET I
C           VERS SA PROPRE CLASSE ET UNE AUTRE CLASSE
C      AMAX=MAXIMUM DE D PARMI LES N OBJETS I
C      B(I)=NUMERO DE LA CLASSE VERS LAQUELLE LE TRANSFERT DE I
C                EST LE PLUS AVANTAGEUX
C      MM=NUMERO DE CLASSE RECEPTRICE
C      LL=NUMERO DE CLASSE EMETRICE
C      JA=LISTE DES ELEMENTS D UN GROUPE POUR L EXTRACTION

       EQUIVALENCE (V(1,1),A(1,1))
       INTEGER P,B
       REAL  MIN
       CALL PLACMT
       NB=2
       IH=10
       IN=NI

       IPOM=0
       IEEUR=0
       LO=II*NI

       READ 114 ,P,KO

   C  LECTURE DU CLASSEMENT INITIAL
         READ 103,(IA(I),I=1,NI)

         READ 106,(MOD(I),I=1,12)
   C  LECTURE ET SYMETRISATION DE LA MATRICE S
         DO 3 N=1,NN
         ID=(N-1)*LO
         K=(N-1)*II
         DO 4 I=1,II
         K1=K+I
         READ MOD,(S(I,J),J=K1,NI)
         DO 5 J=K1,NI
   C     FORMULE VARIANT AVEC LE CONTENU DES CARTES
       5 S(I,J)=(S(I,J)-13.5)/27
       4 S(I,K1)=0
         DO 6 I=2,II
         K1=K+I
         II=I-1
         DO 6 J=1,II
         K2=K+J
       6 S(I,K2)=S(J,K1)
         CALL DWR (16,ID,LO,S)
         CALL DCHK
       3 CONTINUE
         PRINT 107
         DO 7 N=1,NN
         ID1=(N-1)*LO
         CALL DRD (16,ID1,LO,S)
         CALL DCHK
         N1=N-1
         K3=N1*II
         IF (N1) 8,8,9
       9 DO 10 IP=1,N1
         ID2=(IP-1)*LO
```

```
      CALL DRD(16,ID2,LO,A)
      CALL DCHK
      K4=(IP-1)*II
      DO 10 I=1,II
      K2=K3+I
      DO 10 J=1,II
      K1=K4+J
      S(I,K1)=A(J,K2)
   10 CONTINUE
    8 CALL DWR(15,ID1,LO,S)
C  IMPRESSION DE NB BLOCS (NB*II LIGNES)DE LA MATRICE DE SIMILARITE
C       NB=0  PAS D EXTRACTION
C       NB=NN  EXTRACTION TOTALE
      IF(N-NB) 17,17,7
   17 DO 11 I=1,II
      K=K3+I
      PRINT 108 ,K,(S(I,J),J=1,NI)
   11 CONTINUE
      CALL DCHK
    7 CONTINUE
      DO 90 N=1,NN
      ID=(N-1)*LO
      CALL DRD (15,ID,LO,S)
      CALL DCHK
      DO 91 I=1,II
      DO 91 J=1,NI
   91 V(J,I)=S(I,J)
      CALL DWR(15,ID,LO,V)
      CALL DCHK
   90 CONTINUE
C  CALCUL INITIAL DES ATTRACTIONS
   12 DO 13 N=1,NI
      DO 13 J=1,MP
   13 C(N,J)=0
      DO 14 N=1,NN
      ID=(N-1)*LO
      CALL DRD (15,ID,LO,V)
      CALL DCHK
      K1=(N-1)*II
      DO 14 I=1,II
      K=K1+I
      L=IA(K)
      DO 14 J=1,NI
      C(J,L)=C(J,L)+V(J,I)
   14 CONTINUE
C  CALCUL DE LA VALEUR INITIALE DE VL
      VL=0
      DO 16 I=1,NI
      L=IA(I)
   16 VL=VL+C(I,L)
      PRINT 111,VL
      GO=(-NI*(NI-1))/2

C  ****TRAITEMENT****
C
C  MISE A L ETAT INITIAL
   21 ITRANS=0
      IHD=IDUREE(1)
```

```
      PRINT 115,IHD
      DO 26 I=1,NI
   26 B(I)=0
      LL=0
      MM=0
C  RECHERCHE DU TRANSFERT LE PLUS AVANTAGEUX
   28 AMAX=GO
      DO 30 I=1,NI
      U=GO
      L=IA(I)
      IF (B(I)-LL) 81,80,81
   81 IF(B(I)-MM) 31,80,31
   80 CONTINUE
      DO 32 J=1,P
      IF(J.EQ.L)GOTO 32
      IF(C(I,J)-U)32,32,33
   33 U=C(I,J)
      M=J
   32 CONTINUE
      GO TO 36
   31 M=B(I)
      U=C(I,M)
      IF(L.EQ.LL) GO TO 34
      IF(C(I,LL)-U)34,34,35
   35 U=C(I,LL)
      M=LL
   34 IF(L  Q.MM) GO TO 36
      IF(C(I,MM)-U) 36,36,37
   37 U=C(I,MM)
      M=MM
   36 D=C(I,M)-C(I,L)
      B(I)=M
      IF(D-AMAX)30,30,38
   38 AMAX=D
      K=I
   30 CONTINUE
      ITRANS=ITRANS+1
      IF (AMAX)60,60,39
C  REALISATION DU TRANSFERT
   39 LL=IA(K)
      MM=B(K)
      IA(K)=MM
      ID=(K-1)*NI
      CALL DRD(15,ID,NI,V)
      CALL DCHK
      DO 41 I=1,NI
      C(I,LL)=C(I,LL)-V(I,1)
   41 C(I,MM)=C(I,MM)+V(I,1)
      VL=VL+AMAX*2
   42 IF (ITRANS-KO) 43,76,76
   43 IHD=IDUREE(1)
      IF(IH-IHD)76,76,28
   76 IEEUR=1
C  IMPRESSION DU CLASSEMENT SI=
C     -IL N Y A PLUS DE TRANSFERT AVANTAGEUX EN P CLASSES
C     -LES CONDITIONS DE TEMPS ET DE TRANSFERTS SONT ATTEINTES
   60 PRINT 119,ITRANS,VL,AMAX
      DO 61 J=1,P
      KC=0
```

```
      DO 62 I=1,NI
      IF(IA(I)-J)62,63,62
   63 KC=KC+1
      JA(KC)=I
   62 CONTINUE
      IF(KC)61,61,64
   64 PRINT 117,J,KC,(JA(KL),KL=1,KC)
   61 CONTINUE

C  PEUT-ON AMELIORER VL EN AUGMENTANT P
      K=0
      MIN=0
      VL=0
      DO 65 I=1,NI
      J=IA(I)
      VL=VL+C(I,J)
      IF(C(I,J)-MIN)66,65,65
   66 MIN=C(I,J)
      K=I
   65 CONTINUE
      PRINT 120 ,VL,K,MIN
      IF(K)68,68,67
   67 IF(IEEUR.EQ.1)GO TO 68
      IF(P-MP)69,68,68
C  PROGRESSION DE P
   69 P=P+DP

      AMAX=-MIN
      DO 99 J=1,NI
   99 B(J)=P
      GO TO 39
C
C  ****FIN DE TRAITEMENT****
C
C  LE TRAITEMENT S'ARRETE SI=
C     - VL NE PEUT PAS ETRE AMELIORER
C     -LE NOMBRE DE CLASSES EST MAXIMUM
C     - LE NOMBRE DE TRANSFERTS EST MAXIMUM
C     - LE TEMPS IMPARTI AU TRAITEMENT EST DEPASSE
C
C  IMPRESSION ET STOCKGE RUBAN DES RESULTATS

      DO 70 I=1,NI
      PRINT 123 I,(C(I,J),J=1,P)
   70 CONTINUE
      IF(IPOM-1)71,72,71
   72 REWIND 6
      DO 73 N=1,NN
      ID=(N-1)*LO
      CALL DRD (15,ID,LO,V)
      CALL DCHK
      WRITE (6),((V(J,I),I=1,II),J=1,NI)
   73 CONTINUE
      WRITE(6),((C(I,J),J=1,P),I=1,NI)
      WRITE(6),VL,(IA(I),I=1,NI),GO,AMAX,L,M,K
      REWIND 6
      PRINT 105 ((I,IA(I)),I=1,NI)
   71 READ 100,PREM,CHAIN,NBANDE,(TEXTE(K),K=1,3),ITEST
      IF(PREM.NE.3HFIN) GO TO 75
C  ARRET SI PAS D AUTRE TRAITEMENTS
   23 STOP

      END
```

Nous allons présenter maintenant la partie principale des programmes relatifs à l'algorithme "lexicographique" (HIE 01, HIE 02 et HIE 04). Ces programmes ont été composés par Monsieur P. Achard.

HIE 01 –PREORDON– Ce programme lit la matrice d'incidence des données Objets x Attributs. Les variables sont les suivantes :

NB : nombre d'objets (valeur maximum : 150)

LG : $\frac{\text{nombre d'attributs} - 1}{45}$ (valeur maximum : 35)

On désigne dans le programme par

IRM : le nombre de paires d'objets $\left(\frac{NB(NB-1)}{2}\right)$

X(M, L) : la description de l'objet M ; c'est-à-dire un vecteur binaire ayant autant de dimensions que d'attributs ; ce vecteur sera représenté par un nombre dans le système octal. L désigne le numéro du mot correspondant à cette description.

S(M, N) : le nombre d'attributs communs aux objets M et N, l'objet N étant rangé après M dans le tableau (N > M).

S(M, M) : le nombre d'attributs possédés par l'objet M.

S(N, M) : la similarité entre M et N où similarité signifie ici le nombre.

$$\frac{S(M,N)}{S(M,M) + S(N,N) - S(M,N)}$$

(Il n'est pas difficile de modifier dans ce programme les instructions définissant la forme de la similarité choisie).

ISORT : variable d'édition de la préordonnance.

Ce programme calcule, pour tout M, tout N, (N > M), à l'aide du sous-programme "CARDINAL", la similarité S(N, M). Ensuite il fait appel à un sous-programme ORCLAS pour ranger les couples d'objets (M, N) dans l'ordre de leur similarité décroissante. Le programme édite le résultat de ce tri et le transcrit sur ruban. Il fait appel ensuite au sous-programme ARBRE (dont la présentation va suivre).

HIE 01

```
      PROGRAM PREORDON
      INTEGER R
      EQUIVALENCE(R,S)
      INTEGER PRESENT
      INTEGER X,S,U,CARDINAL

      INTRODUCTION DES DONNEES


      READ 3,NB,LG
      IRM=NB*(NB-1)/2
      DO2 M=1,NB
    2 READ(15,1) (X(M,L),L=1,LG)

      CALCUL DE S, S+U, S+V


      DO 4 M=1,NB
      MP1=M+1
      DO 5 N=MP1,NB
      DO 6L=1,LG
    6 INT(L)=X(M,L).AND.X(N,L)
    5 S(M,N)=CARDINAL(INT,LG)
      DO 7L=1,LG
```

```
  7 U(L)=X(M,L)
    PRINT 10,M,(U(L),L=1,LG)
  4 S(M,M)=CARDINAL(U,LG)

    CALCUL  DE  LA  SIMILARITE


    NLI=0
    DO12 M=1,NB
    MP1=M+1
    DO 12 N=MP1,NB
    NLI=NLI+1
    INDICE(1,NLI)=N + 150*(M-1)
    IF (S(M,N))711,711,712
711 S(N,M)=0 $ GO TO 12
712 CONTINUE
    HAUT =S(M,N)$BAS=S(M,M)+S(N,N)-S(M,N)$RAP=HAUT/BAS
    S(N,M)=1000000*RAP
 12 CONTINUE

    TRI  CLASSEMENT


    CALL ORCLAS

    CALCUL  DES  RANGS  ET  STOCKAGE  SUR  BANDE


    IDECAL=0


    MP=(INDICE(1,1)-1)/150+1
    NP=INDICE(1,1)-(150*(MP-1))

    DO 51 IC=1,IRM

    LI=LI+1 IDECAL=IDECAL+1 M=MP N=NP
    WRITE(10)ISELEC,M,N
    MP=(INDICE(1,IC+1)-1)/150+1
    NP=INDICE(1,IC+1)-150*(MP-1)
    IF(S(N,M)-S(NP,MP))920,511,512
511 IF(LI-12)51,513,513
513 PRINT24,(ISORTM(LO),ISORTN(LO),LO=1,12)$LI=0$GO TO 51
512 PRINT24,(ISORTM(LO),ISORTN(LO),LO=1,LI)
    ISELEC=ISELEC+IDECAL$IDECAL=0$LI=0$PRINT23,ISELEC
    GO TO 51
920 IF(IC.NE.IRM)GO TO 921
 51 CONTINUE
999 CONTINUE
    ENDFILE 10
    REWIND 10

    APPEL AU SOUS-PROGRAMME HIERARCHIE DE CLASSIFICATIONS


    CALL ARBRE
    GO TO 888
921 PRINT 543
543 FORMAT(10X,*ERREUR DE PROCEDURE *)
888 CONTINUE
    END
```

EXEMPLE D'EDITION DE LA PREORDONNANCE

```
COUPLES DE RANG     185
    ( 70, 87)(  2, 89)(

COUPLES DE RANG     187
    (  2, 28)(

COUPLES DE RANG     188
    ( 13, 72)(

COUPLES DE RANG     189
    ( 74, 89)(

COUPLES DE RANG     190
    ( 50, 70)( 28, 92)( 11, 20)(

COUPLES DE RANG     193
    ( 50, 57)(

COUPLES DE RANG     194
    (  4, 76)(

COUPLES DE RANG     195
    ( 11, 70)(

COUPLES DE RANG     196
    ( 11, 87)(

COUPLES DE RANG     197
    ( 65, 87)(

COUPLES DE RANG     198
    ( 20, 92)(

COUPLES DE RANG     199
    ( 21, 94)( 20, 70)( 20, 23)(  2, 57)(

COUPLES DE RANG     203
    (  2, 67)(

COUPLES DE RANG     204
    (  2, 50)(

COUPLES DE RANG     205
    ( 57, 72)(

COUPLES DE RANG     206
    ( 28, 50)(

COUPLES DE RANG     207
    ( 11, 67)(

COUPLES DE RANG     208
    ( 50, 92)(

COUPLES DE RANG     209
    (  7, 48)(

COUPLES DE RANG     210
    ( 65, 89)(

COUPLES DE RANG     211
    ( 57, 89)(
```

HIE 02 – CARDINAL

Ce programme dénombre le nombre de propriétés possédées par un objet quelconque. Pour ceci il utilise un masque mobile.

Signification des variables

K (L) : Description de l'objet

I C : Masque

Ci-dessous : listing de la fonction.

```
      FUNCTION CARDINAL(K,LG)
      INTEGER CARDINAL
      DIMENSION K(35)
      CARDINAL=0
      DO1L=1,LG
      IC=1
      DO 1 N=1,45
      IN=K(L).AND.IC
      IC=2*IC
    1 IF(IN.NE.0)CARDINAL=CARDINAL+1
      RETURN
      END
```

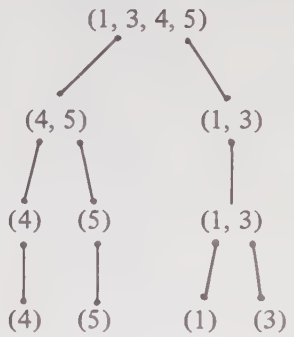
HIE 04 – ARBRE

Ce programme peut indifféremment être appelé à partir de PREORDON ou être exécuté indépendamment.

Il a pour entrée une bande magnétique sur laquelle est inscrit en binaire une suite de triplet (rang de la paire d'objet, numéro du premier élément de la paire, numéro du deuxième élément de la paire). Cette suite est ordonnée par rang croissant des paires d'objets (correspondant au pré-ordre donné sur l'ensemble des paires d'objets).

Il a pour but de construire une hiérarchie de classification selon l'algorithme "lexicographique".

La chaîne de partitions obtenue est représentée en utilisant la notation polonaise. Les nombres négatifs sont utilisés pour représenter les regroupements, les nombres positifs désignent les numéros des objets. Ainsi la hiérarchie suivante sur l'ensemble des objets 1, 3, 4, 5



sera représentée par le mot – 3 – 2 4 5 – 1 3 1,

Signification des symboles de variables utilisés dans le programme :

ISELEC : rang de la paire

M, N : objets d'une paire

NIVO : étiquette de la parenthèse

GROUPE : représentation polonaise.

Ci-dessous : listing du programme.

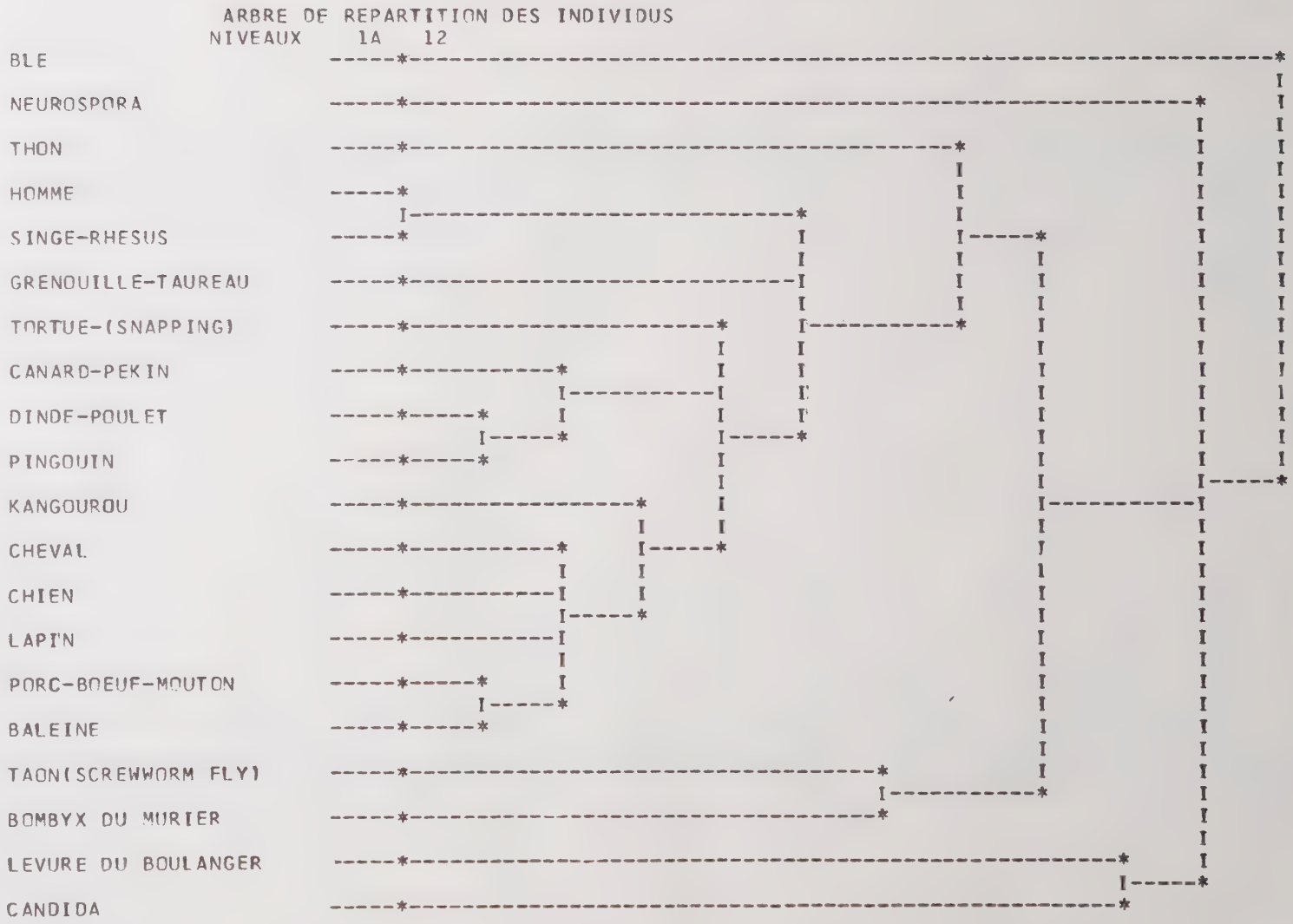
HIE 04

```
      SUBROUTINE ARBRE
       COMMON/A/GROUPE
      INTEGER TRANS
      INTEGER GROUPE
      LIP=0
      IMEM=1
      NIVO=1
      DO11I=1,300
    1 GROUPE(I)=0
    2 CONTINUE
      READ (  10)ISELEC,M,N
      IF(EOF,10)69,33
   33 CONTINUE
      IF(ISELEC-IMEM)3,4,5
    3 PRINT 6

      GO TO 999
    6 FORMAT(*  ERREUR SUR LA BANDE 10 -DECLASSEMENT*)
    5 NIVO=NIVO+LIP
      LIP=0
      IMEM=ISELEC
    4 CONTINUE
      NOMBRE=1
      IGRO(1)=IGRO(2)=0
      J=0
      K=1
      DO27L=1,150
   27 IRAN(L)=1
       DO67I=1,300
      NOMBRE=NOMBRE-1
      IF(GROUPE(I))11,12,13
   11 IF(NOMBRE .LT.0)NOMBRE=0
      NOMBRE=NOMBRE+2
      GO TO 67
   12 CONTINUE
      IF(NOMBRE.GE.0.AND.I.NE.1)GO TO 68
      IF(M.EQ.N) GO TO 9
    8 GROUPE(I)=-NIVO
      GROUPE(I+1)=M
        GROUPE(I+2)=N
      LIP=1
      GO TO  2
    9 CONTINUE
      IF(IGRO(2).NE.0)GO TO 68
      NI=IGRO(1)
      IAMPL=IRAN(NI)$IREST=I-IAMPL
      DO 20 ID=1,IREST
   20 GROUPE(I-ID+2)=GROUPE(I-ID)
      GROUPE(IAMPL)=-NIVO$GROUPE(IAMPL+1)=M
      LIP=1
      GO TO 2
   13 CONTINUE
      IF(GROUPE(I).NE.M.AND.GROUPE(I).NE.N)GO TO 65
      NM=(N*M)/GROUPE(I)
      N=M=NM
      J=J+1
      IF(J.GT.2)GO TO 68
```

```
      IGRO(J)=K
   65 IF(NOMBRE)68,50,67
   50 K=K+1
      IRAN(K)=I+1
      IF(J.LT.2)GO TO 67
      IF(IGRO(2)-IGRO(1)-1)51,52,53
   51 GO TO 2
   52 LU=IGRO(1)
      LIMAX=300-IRAN(LU)
      DO54LI=1,LIMAX
      LILI=300-LI
      GROUPE(LILI+1)=GROUPE(LILI)
   54 CONTINUE
      GROUPE(300 -LIMAX)=-NIVO
      LIP=1
      GO TO 2
   53 MUS=IGRO(1)+1
      LI=IRAN(MUS)
      MUSUP=IGRO(2)
      LISUP=IRAN(MUSUP)-1

      MU0=IGRO(1)
      DO555IND=LI,LISUP
      TRANS(IND)=GROUPE(IND)
  555 CONTINUE
      IAMPL=IRAN(MUSUP+1)-IRAN(MUSUP)
      IAMP0=IRAN(MUS)-IRAN(MU0)
      LI0=IRAN(MU0)
      DO545JK=1,IAMP0
  545 GROUPE(LI-JK+1)=GROUPE(LI-JK)
      GROUPE(LI0)=-NIVO
      DO544 JL=1,IAMPL
  544 GROUPE(LI+JL)=GROUPE(LISUP+JL)
      LIMIT=300-IAMPL-LISUP
      DO454JM=1,LIMIT
  454 GROUPE(300-JM+1)=GROUPE(300-JM)
      LIMT=LISUP-LI+1
      DO445JN=1,LIMT
  445 GROUPE(LI+IAMPL+JN)=TRANS(LI+JN-1)
      LIP=1
      GO TO 2
   67 CONTINUE
   68 CONTINUE
      PRINT71
   71 FORMAT (*   ERREUR DANS LA PROCEDURE * )
   69 CONTINUE
       PRINT72,GROUPE
   72 FORMAT(*     REPRESENTATION POLONAISE DE L'ARBRE*/(5X,2016))
  999 CONTINUE
      RETURN
      END
```

Il nous reste encore à exprimer, dans le cadre de cette annexe, les programmes relatifs au problème de la h-classificabilité étudié dans la deuxième partie du chapitre 4. Ces programmes sont intitulés, HC0, HC1 et HC2. HC0 permet d'établir la distribution observée, $D(\omega)$, sur l'exemple réel qu'on traite. HC1 détermine la distribution théorique $D(\omega)$, dans l'hypothèse N et le cas particulier spécifié ci-dessus (Ch. 4 § II 3.1), à partir du calcul des valeurs $M(m)$ et $P(m)$; enfin HC2 établit une bonne approximation de cette même distribution théorique à partir des valeurs $M'(m)$.

EXEMPLE D'EDITION DE L'ARBRE



Nous-mêmes avons mis au point ces programmes ; toutefois l'organisation générale de HC0, dont l'entrée dépend de la sortie du programme PREORDON, nous a été fournie par P. Achard.

HC0 — L'entrée de la préordonnance se fait séquentiellement ; chaque élément de la séquence est un tri-uple d'entiers (IRAN, M, N) où IRAN indique le rang de la paire $\{M, N\}$. Les rangs $(R(M, N), M < N)$ des NB(NB − 1)/2 paires différentes sont stockés dans la partie supérieure (au-dessus de la diagonale principale) d'un tableau de dimension NB × NB. La partie inférieure du tableau (au-dessous de la diagonale principale) contiendra les entiers $(R(N, M), M < N)$ où $R(N, M)$ est le rang de la classe consécutive à celle de la paire $\{M, N\}$, et dont l'intérêt sera bientôt mis en évidence.

Un tableau à une ligne DIST de dimension NB(NB − 1)/2 sera supposé indexé sur l'ensemble F des paires de manière conforme à la préordonnance et contiendra la distribution $D(\omega)$. Considérons la kième position de ce tableau supposé rempli :

Si DIST(K) est strictement inférieur à 1, DIST(K) est la proportion des triplets pour lesquelles la paire médiane et la paire supérieure enserrent la paire correspondante à la position K. Si DIST(K) est supérieur à 1, la position K correspond à un début de classe de la préordonnance, le rang correspondant étant la partie entière de DIST(K) et le reste

$$\text{DIST}(K) - [\text{DIST}(K)]$$

à la même signification que ci-dessus.

La distribution $D(\omega)$ est finalement recueillie dans les tableaux MASSOB et PREMP.

Dans ce programme, l'ensemble des triplets est passé en revue ; si (I, J, K) est un triplet donné pour lequel, par exemple, on a

$$R(J, K) \leqslant R(I, J) \leqslant R(I, K)$$

avec

$$I < J < K,$$

les positions du tableau DIST, correspondant à l'intervalle ouvert]Méd. Sup.[, sont définies par l'intervalle dans N $[R(J, I), (R(I, K) - 1)]$.

HC1-HC2 — Il n'y a rien de particulier à signaler sur ces programmes sinon qu'il s'est avéré commode d'introduire au début de chacun d'eux une table de Log(n !), $n = 1, 2, \ldots, 300$. Dans HC1 et HC2, $M(m)$ est représenté par MASS(M), $M'(m)$ par MASS2(M) et $P(m)$ par PROB(M).

HC0

DEFINITION DES PARAMETRES

```
      EQUIVALENCE (R(1),RR(1))
      EQUIVALENCE(DIST(1),MASSOB(1))
      READ(5,1) NB
      X=NB*(NB-1)*(NB-2)
      TRIPL=X/6.0
      UNIT=1.0/TRIPL
      PAIR=NB*(NB-1)/2
      UNIPR=1./PAIR
      NPAIR=PAIR
```

ENTREE DE LA PREORDONNANCE ET CALCUL DE $R(M, N)$ ET DE $R(N, M)$ POUR TOUT COUPLE (M, N), $(M < N)$.

```
      READ (5,2) IRAN,M,N
      R(M,N)=IRAN
      RAN=IRAN
      DIST(IRAN)=RAN
      R(N,M)=0
      IPREC=IRAN
    5 N1M1=N+50*(M-1)
      READ(5,2,END=9) IRAN,M,N
      R(M,N)=IRAN
      RAN=IRAN
      DIST(IRAN)=RAN
      IF(IRAN.GT.IPREC) GO TO 7
      R(N,M)=N1M1
      GO TO 5
    7 R(N,M)=0
      IPREC=IRAN
      IRET=RR(N1M1)
      RR(N1M1)=IRAN
    8 IF(IRET.EQ.0) GO TO 5
      NMRET=IRET
      IRET=RR(NMRET)
      RR(NMRET)=IRAN
      GO TO 8
    9 CONTINUE
```

DETERMINATION POUR CHAQUE TRIPLET (I, J, K) DE LA PAIRE MEDIANE ET DE LA PAIRE SUPERIEURE.

```
      NBM2=NB-2
      DO 100 I=1,NBM2
      I1=I+1
      NBM1=NB-1
      DO 99 J=I1,NBM1
      J1=J+1
      DO 98 K=J1,NB
      IF(R(I,J)-R(J,K)) 110,115,120
  110 IF(R(I,K)-R(I,J)) 25,25,30
   25 SUP=R(J,K)
      MED=R(J,I)
      GO TO 97
   30 IF(R(I,K)-R(J,K)) 35,98,45
   35 SUP=R(J,K)
      MED=R(K,I)
      GO TO 97
   45 SUP=R(I,K)
      MED=R(K,J)
      GO TO 97
  115 IF(R(I,K)-R(I,J)) 98,98,55
   55 SUP=R(I,K)
      MED=R(J,I)
      GO TO 97
  120 IF(R(I,K)-R(J,K)) 60,60,65
   60 SUP=R(I,J)
      MED=R(K,J)
      GO TO 97
   65 IF(R(I,K)-R(I,J)) 70,98,80
   70 SUP=R(I,J)
      MED=R(K,I)
      GO TO 97
   80 SUP=R(I,K)
      MED=R(J,I)
   97 IF(MED.EQ.SUP) GO TO 98
```

CHARGEMENT DU TABLEAU DIST

```
      ISUP=SUP-1
      DO 13 LL=MED,ISUP
   13 DIST(LL)=DIST(LL)+UNIT
   98 CONTINUE
   99 CONTINUE
  100 CONTINUE
```

LA DISTRIBUTION CHERCHEE

```
      M=0
      DO 150 KK=1,NPAIR
      IF(DIST(KK)-1.)51,52,52
   51 PREMP(M)=PREMP(M)+UNIPR
      GO TO 150
   52 M=M+1
      DRET=DIST(KK)
      IENT=INT(DRET)
      ENT=IENT
      MASSOB(M)=DIST(KK)-ENT
      PREMP(M)=UNIPR
  150 CONTINUE
      STOP
      END
```

HC1

INTRODUCTION DE LA TABLE

```
      TABL(1)=0.
      DO 10 N=2,301
      X=N-1
   10 TABL(N)=TABL(N-1)+ALOG(X)
```

INTRODUCTION DES DONNEES IT, K,

```
      KM1=K-1
      KP1=K+1
      JT=IT+1
      JTMK=JT-K
      TT=TABL(JT)
      C=ALOG(3.)+3*(TABL(KP1)+TABL(JTMK))-2*TT
      D=2*TABL(KP1)+2*TABL(JTMK)-TT
      DO 999 M=1,KM1
```

CALCUL DE MASS(*M*) ET PROB(*M*)

```
      MP1=MAXO(M+1,2*K-IT)
      DO 100 IR=MP1,K
      DO 100 JQ=1,M
      IQ=JQ-1
      DO 100 JP=1,M
      IP=JP-1
      JR=IR+1
      IF((IP+IQ.GT.IR+K).OR.(IQ+IR.GT.IP+K).OR.(IR+IP.GT.IQ+K))GO TO100
      LSUP=MAXO(0,IP+IQ-K,IQ+IR-K,IR+IP-K)
      LINF=MINO(IT+IP+IQ+IR-3*K,IP,IQ)
      IF(LINF-LSUP)100,45,45
   45 SIG=0.
      JLSUP=LSUP+1
      JLINF=LINF+1
      DO 55 JL=JLSUP,JLINF
      L=JL-1
      U=TABL(JL)+TABL(JP-L)+TABL(JQ-L)+TABL(JR-L)
      V=TABL(K+JL-IP-IR)+TABL(K+JL-IQ-IR)+TABL(K+JL-IP-IQ)
      W=TABL(JT+IP+IQ+IR-3*K-L)
      REL=C-(U+V+W)
      EL=EXP(REL)
   55 SIG=SIG+EL
      MASS(M)=MASS(M)+SIG
  100 CONTINUE
      Y=TABL(MP1)+2*TABL(KP1-M)+TABL(JT+M-2*K)
      RCP=D-Y
      PROB(M)=EXP(RCP)
```

CALCUL DES SOMMES ET DE *H*

```
      SOMAS=SOMAS+MASS(M)
      SPROB=SPROB+PROB(M)
      H=H+PROB(M)*MASS(M)
      IF(PROB(M)*MASS(M).LT.H*1.E-3) GO TO 66
  999 CONTINUE
   66 WRITE(6,3) SOMAS,SPROB,H
      STOP
      END
```

HC2

INTRODUCTION DE LA TABLE

```
      TABL(1)=0.
      DO 10 N=2,301
      X=N-1
   10 TABL(N)=TABL(N-1)+ALOG(X)
```

INTRODUCTION DES DONNEES IT, K,

```
      KM1=K-1
      KP1=K+1
      JT=IT+1
      JTMK=JT-K
      TT=TABL(JT)
      TAK=TABL(KP1)
      TTMK=TABL(JTMK)
      U=IT
      V=K
      C2=ALOG(3.)+4*(TAK+TTMK)-2*TT
      D=2*TAK+2*TTMK-TT
      DO 999 M=1,KM1
```

CALCUL DE MASS2(M) et PROB(M)

```
      MP1=MAXO(M+1,2*K-IT)
      DO 100 IR=MP1,K
      DO 100 JQ=1,M
      IQ=JQ-1
      DO 100 JP=1,M
      IP=JP-1
      JR=IR+1
      IF((IP+IQ.GT.IR+K).OR.(IQ+IR.GT.IP+K).OR.(IR+IP.GT.IQ+K))GO TO100
      P=IP
      Q=IQ
      R=IR
      UMV=U-V
      PAR=P*R/V+(V-P)*(V-R)/UMV
      ATRM=C2-TABL(JP)-2*TABL(KP1-IP)-TABL(JT+IP-2*K)-TABL(JR)
     )-2*TABL(KP1-IR)-TABL(JT+IR-2*K)
      IF(PAR.EQ.0) GO TO 50
      ATRM=ATRM-TABL(JQ)-PAR+Q*ALOG(PAR)
      TRM=EXP(ATRM)
      GO TO 99
   50 TRM=EXP(ATRM)
   99 MASS2(M)=MASS2(M)+TRM
  100 CONTINUE
      Y=TABL(MP1)+2*TABL(KP1-M)+TABL(JT+M-2*K)
      RCP=D-Y
      PROB(M)=EXP(RCP)
```

CALCUL DES SOMMES ET DE H2

```
      SOMAS2=SOMAS2+MASS2(M)
      SPROB=SPROB+PROB(M)
      H2=H2+PROB(M)*MASS2(M)
      IF(PROB(M)*MASS2(M).LT.H2*1.E-3) GO TO 66
  999 CONTINUE
   66 WRITE(6,32)SOMAS2,SPROB,H2
      STOP
      END
```

ANNEXE II

EXEMPLES RÉELS

Ici nous rendrons compte du traitement de quelques expériences. Le premier exemple illustre la méthode de S. Régnier, les deux autres, certains aspects de la nôtre.

1. *UN EXEMPLE DE GEOGRAPHIE ECONOMIQUE*

par S. Régnier

Une recherche menée par Mr. Guillot de l'Ecole Pratique des Hautes Etudes sur "Le prix du blé au dix-neuvième siècle" a conduit à l'analyse de l'évolution de ce prix dans chaque département au courant des années "1825 à 1873".

Des études préalables autorisent à penser que certains départements "pôles" orientent l'évolution du prix dans leurs zones d'influence respectives. Le but de l'étude est d'essayer de déterminer ces zones.

Il n'est pas exclu qu'une zone d'influence puisse s'étendre à des départements non limitrophes les uns des autres et comporte de ce fait plusieurs parties connexes.

Le département i peut être représenté par la série de prix moyens annuels.

$p_i^t \qquad t = 1, 2, \ldots, 49$ années

$i = 1, 2, \ldots, 89$ départements.

La méthode utilisée est celle définie par l'algorithme des transferts, [cf. Ch. 3 § II], qui suppose définie une mesure de ressemblance sur l'ensemble des couples d'objets. La première question posée est donc : quelle mesure de similarité adopter entre départements ?

La ressemblance entre deux séries p_i et p_j peut être mesurée par un coefficient de corrélation

a) linéaire

b) logarithmique

c) entre rangs

Il s'agit chaque fois d'un coefficient sans dimension compris entre -1 et $+1$, et qui vaut 1 seulement lorsque les prix, leur logarithme, ou leur rang relatifs dans le temps sont liés par une relation linéaire soit :

a) $p_i^t = a + bp_j^t$

b) $p_i^t = a(p_j^t)^b$

c) $r_i^t = a + br_j^t$ (r_i^t = rang du prix p_i^t dans la suite des 49, $t = 1, 2, \ldots$ ou 49)

a) signifie d'ailleurs que les accroissements de prix d'une année à une autre dans les départements i et j seraient proportionnels.

b) que les élasticités $dp_i/p_i, dp_j/p_j$ seraient proportionnelles.

Divers arguments militent en faveur de ces 3 options. On a convenu de les étudier toutes les trois.

Nous avons d'abord calculé les prix moyens annuels et les corrélations linéaires et logarithmique des 89 départements deux à deux, soit R_i^j.

Il est apparu que toutes ces corrélations sont fortes et > 0.5. Cela signifie que tout le processus d'évolution des 89 prix est dominé par un trend général à

PAS-DE-CALAIS 3
NORD 3
SOMME 3
SEINE INFERIEURE 3
AISNE 3
ARDENNES 3
MANCHE 2
CALVADOS 2
OISE 3
EURE 3
SEINE-ET-OISE 3
MARNE 3
MEUSE 12
MOSELLE 12
BAS RHIN 12
FINISTÈRE 1
CÔTES DU NORD 2
ORNE 3
EURE-ET-LOIR 3
SEINE-ET-MARNE 3
MEURTHE-ET-MOSELLE 12
MORBIHAN 1
ILLE ET-VILAINE 2
MAYENNES 2
SARTHE 3
AUBE 3
HAUTE MARNE 12
VOSGES 12
HAUT RHIN 12
LOIR-ET-CHER 3
LOIRET 3
YONNE 3
LOIRE-INFÉRIEURE 10
MAINE-ET-LOIRE 9
INDRE-ET-LOIRE 9
CÔTE-D'OR 8
HAUTE-SAÔNE 12
DOUBS 8
VENDÉE 9
DEUX-SÈVRES 9
VIENNE 9
INDRE 7
CHER 7
NIÈVRE 7
SAÔNE-ET-LOIRE 8
JURA 8
CHARENTE-MARITIME 4
CHARENTE 9
HAUTE VIENNE 4
CREUSE 11
ALLIER 11
PUY DE DÔME 11
LOIRE 8
RHÔNE 8
AIN 8
HAUTE-SAVOIE 8
SAVOIE 8
GIRONDE 4
DORDOGNE 4
CORRÈZE 6
CANTAL 5
HAUTE-LOIRE 5
ISÈRE 8
LOT-ET-GARONNE 4
LOT 4
AVEYRON 4
LOZÈRE 6
ARDÈCHE 6
DRÔME 6
HAUTES-ALPES 6
LANDES 4
TARN-ET-GARONNE 4
GERS 4
TARN 4
GARD 6
VAUCLUSE 6
BASSES ALPES 6
ALPES MARITIMES 6
BASSES-PYRENEES 4
HAUTE GARONNE 4
HERAULT 6
BOUCHES DU RHÔNE 6
VAR 6
HAUTES PYRENEES 4
ARIÈGE 4
AUDE 6
PYRENEES ORIENTALES 6
CORSE 6

l'échelle nationale et que les courbes des prix dans 2 départements sont en gros parallèles.

Dans le cadre de la méthode de classification envisagée, cet état des données conduira à une partition qui groupe tous les départements dans une seule classe.

Il nous fallait donc une mesure de ressemblance qui mette mieux en lumière les différences entre départements, et ne soit pas uniformément supérieure à 1/2.

On aurait pu soustraire des R_{ij} un nombre arbitraire (tel que leur moyenne moins 0.5).

On a préféré envisager un modèle :

$$p_i^t = g^t + M_i + D_i^t$$

ou g^t = tendance nationale l'année t.

M_i = écart propre au département i $(\sum_i M_i = 0)$

D_i^t = suite des 49 vecteurs aléatoires indépendants, de moyenne nulle et matrice de variance V.

Dans ce modèle g^t est estimé par p^t = *moyenne des 89 prix l'année t*, les paramètres de V ne peuvent pas tous être estimés, mais une mesure de dépendance entre les prix p_i^t p_j^t est donnée par la corrélation des séries chronologiques : $q_i^t = p_i^t - p^t$ = écart du prix p_i^t à la moyenne nationale. Ce même modèle peut être appliqué aux logarithmes des prix et prend alors une signification économique assez simple :

p'^t = moyenne logarithmique des prix du blé l'année t constitue un indice des prix pour l'année t.

$q_i'^t$ est le logarithme du prix relatif du département i, un nombre pur indépendant des fluctuations de la valeur de la monnaie.

On a déterminé les deux partitions centrales correspondant respectivement aux écarts bruts (q_i^t) et aux écarts sur les logarithmes $(q_i'^t)$. La classification obtenue à partir des $q_i'^t$ est représentée dans la figure suivante ; on distingue douze classes soit douze "zones d'influence". A une exception près, une même zone est formée de départements contigus ; cette exception concerne la zone numérotée 6 (PROVINCIA ROMANA) où le département de CORREZE est séparé des autres par le CANTAL. La partition,fondée sur les écarts bruts, diffère assez peu de la précédente ; elle comprend dix classes au lieu de douze définies par suppression de la classe de numéro 10, réunion des classes 1 et 2 et quelques petites différences ; notamment pour cette classification c'est la zone 8 (ALPES-JURA) qui présente un département excentrique : ALPES MARITIMES. Signalons que pour chacune des deux partitions, on obtient un coefficient de corrélation entre écarts qui vaut en moyenne 0,7 pour deux départements pris dans la même classe.

Ces résultats ayant donné entière satisfaction au spécialiste, nous n'avons pas poursuivi le calcul à partir des rangs des prix dans le temps.

2. *UN EXEMPLE DE PSYCHO-SOCIOLOGIE*[1]

Le problème de la définition d'une typologie s'est posé de manière cruciale dans le cadre d'une recherche menée par M.J. Chombard de Lauwe dont un aspect est l'étude de l'image de l'enfant dans les "moyens de communication de masse" ; c'est-à-dire la recherche des modèles que les adultes créent et proposent aux enfants dans la littérature dite "pour enfants".

La donnée est un ensemble de 1500 personnages enfantins issus d'un échantillon d'ouvrages de cette littérature édités entre 1880 et 1960. Précisons que l'unité d'échantillonnage est *le texte*, que celui-ci soit un roman, une biographie, un conte ou un journal ; il n'est pas dans notre propos d'exposer ici les critères qui ont prévalu pour le choix de telle ou telle œuvre. 150 attributs ont été retenus permettant de décrire, pour un personnage donné, sa psychologie, le milieu dans lequel il vit, ses activités, ses relations avec autrui etc.

A priori deux voies de travail se sont présentées. La première consiste à définir une classification sur l'ensemble des personnages cherchant par la suite, pour une classe donnée, l'ensemble des attributs les plus dominants dans cette classe, ce dernier définissant un type. La deuxième voie, plus directe, consiste à établir une classification sur l'ensemble des

(1) Nous avons établi ce texte avec l'aide précieuse de Cl. Bellan collaborateur de M.J. Chombard de Lauwe du Centre d'Ethnologie Sociale et de Psychosociologie.

attributs, une classe formée exprimant un type. C'est cette seconde solution qui a été adoptée.

Ici donc, avec les notations du texte, E désigne l'ensemble des attributs de description et A l'ensemble des personnages ; chaque élément de E, (soit chaque item), est représenté par un vecteur logique à 1500 composantes, la i-ème étant égale à 1 si l'attribut est présent chez le i-ème personnage et 0 sinon.

Plutôt que d'établir une seule partition, il est préférable de construire une hiérarchie de classifications, car de la sorte on fait apparaître les types les plus marqués dès les premiers niveaux de la hiérarchie et les autres. Par conséquent la procédure adoptée est celle de l'algorithme "lexicographique" [cf. Ch. 3 § III] dont l'argument est une préordonnance sur E que l'on détermine à partir du choix d'un indicateur de ressemblance [cf. Ch. 1 § III] ; les programmes relatifs à la méthode étant HIE 01, et HIE 02 et HIE 04.

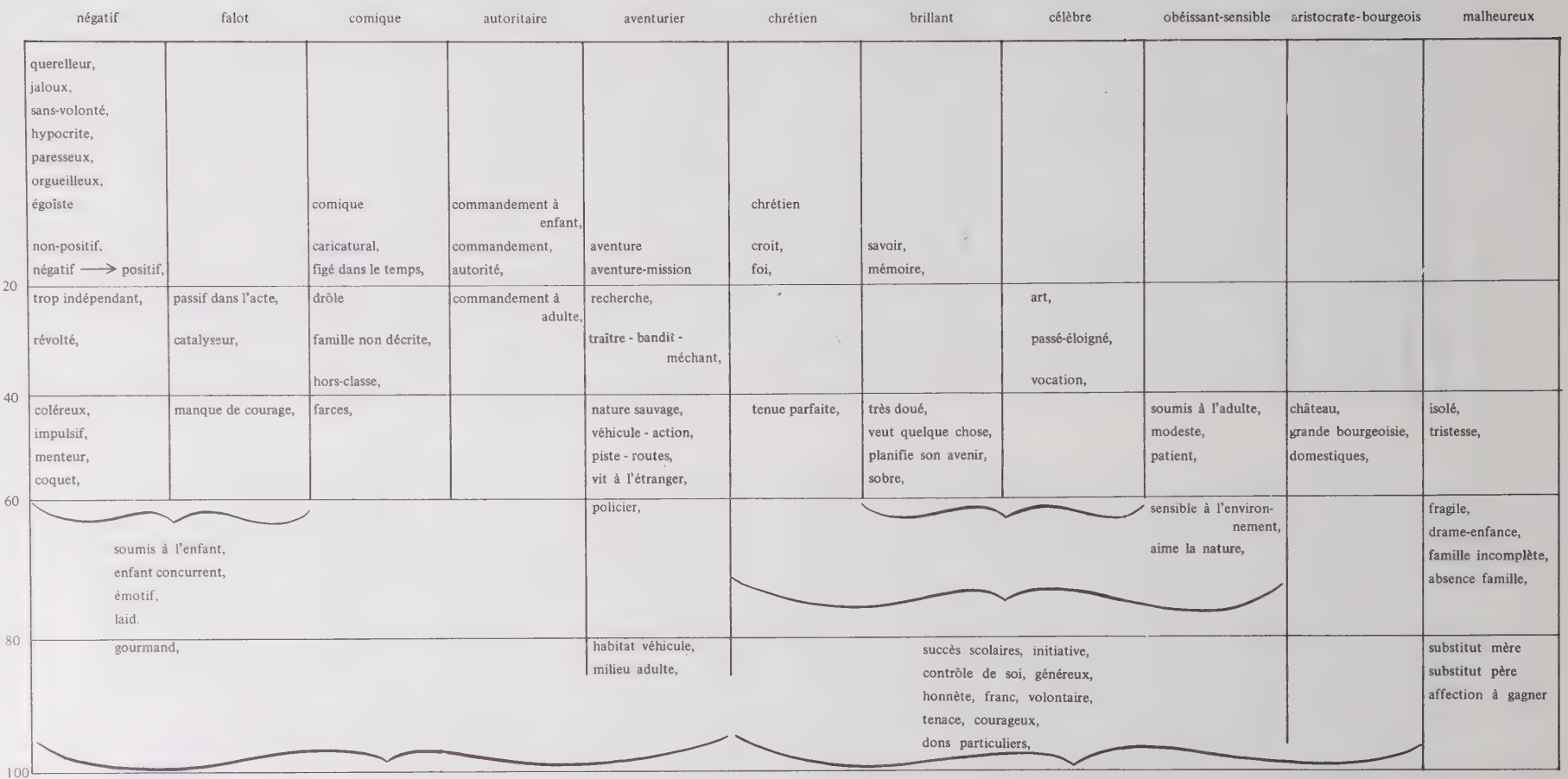
La première question posée est ainsi le choix d'un indice de similarité ; une telle question aurait été sans importance si la variance sur E, de la statistique N_i : nombre d'individus possédant un même attribut, est faible ; malheureusement il n'en est rien comme nous l'a montré le graphique de la distribution de cette statistique. Il est manifeste que dans le choix de l'indice , s : nombre de personnages possédant simultanément deux attributs donnés, doit jouer un rôle important ; en effet, la présence commune, de deux attributs, chez un individu donné, peut être significative de leur ressemblance alors que leur absence commune n'indique rien.

Notre première réaction a été de découvrir sur l'ensemble E des propriétés des parties telles que sur chacune d'elles la distribution de la statistique N_i ait une variance modérée. Seule la classe des attributs, dont la fréquence relative de présence est inférieure à 24/100, a été retenue comme présentant une certaine homogénéité. Cette classe qui contient 98 items sur les 150 étudiés, a été analysée en prenant comme indice $s/s + u + v$: proportion de sujets possédant simultanément deux attributs donnés parmi ceux qui possèdent soit l'un soit l'autre. Les résultats ont été très encourageants ; ils ont permis d'une part, de dégager différents types, certains plus prononcés que d'autres et d'autre part d'écarter quelques attributs secondaires qui n'intervenaient pas du tout dans la formation des classes.

Il nous est resté toutefois à effectuer un traitement global pour lequel l'indice s centré et réduit [cf. Ch. 1 § IV.3] s'impose. Ce traitement, qui a porté sur 125 items, a donné entière satisfaction ; nous allons commenter brièvement les résultats qu'on trouvera exprimés de manière synthétique dans le tableau suivant.

Nous avons pu distinguer onze types différents pour lesquels des étiquettes quelque peu arbitraires ont été portées sur les différentes colonnes du tableau. Dès le vingtième niveau les noyaux de six types différents sont mis à jour : "négatif", "comique", "autoritaire", "aventurier", "chrétien" et "brillant". Le noyau le plus remarquable est celui qui contient la plupart des éléments les plus caractéristiques du type négatif. Ce dernier noyau reste relativement stable durant les vingt niveaux suivants ne s'enrichissant que de deux attributs ; alors que le personnage autoritaire s'achève et que le comique est près de l'être. C'est également à ces niveaux que le type aventurier se précise et celui de l'enfant célèbre apparaît. Au bout du soixantième niveau les types : "négatif", "falot", "comique", "autoritaire", "chrétien", "brillant", "célèbre" et "bourgeois" sont complètement formés, le type "aventurier" étant près de l'être ; d'autre part apparaissent deux groupements relatifs aux types de l'enfant obéissant et de l'enfant malheureux. A partir du soixantième niveau "négatif" et "falot" se rejoignent et s'adjoignent quelques attributs pour former un type négatif assez général. Par ailleurs "brillant" et "célèbre" se réunissent pour donner lieu au type de l'enfant modèle ; à ce dernier s'agrègent "chrétien" et "obéissant", il restera encore à ce type, déjà riche, de récolter une dizaine d'attributs, après le quatre-vingtième niveau, pour constituer un large type normatif. Simultanément les types "aventurier" et "malheureux" achèvent de se préciser. Pour terminer il est intéressant de noter que le "bourgeois" est plus proche du type normatif que du type négatif et que l'inverse a lieu pour le "comique".

CLASSES FORMEES AVANT LE NIVEAU N°



Note. Soulignons que les items qui figurent ici sont exprimés sous une forme très schématique et renvoient à des niveaux d'analyse très différents ; par exemple "traître-bandit-méchant" est relatif à l'environnement du héros plutôt qu'à lui-même.

Le fait que le degré de formation des différents types à un niveau donné de la hiérarchie ne soit pas le même provient de deux facteurs de nature statistique ; le premier est la fréquence d'un type donné dans la littérature étudiée (un type apparaîtra d'autant plus rapidement qu'il est plus fréquent) et le second la richesse du vocabulaire décrivant ce type ; c'est-à-dire le nombre d'attributs intervenant dans sa description.

Il est maintenant permis de se demander quel a été l'apport de ce traitement pour le spécialiste. Si les résultats sont rarement surprenants, il eut tout de même été extrêmement difficile de les prévoir ; voire, de les retrouver, en se fiant seulement à l'intuition et au travail de dépouillement. Par exemple, l'importance relative des caractères dans la constitution d'un type donné était inconnue ; il a même été constaté que certains attributs que l'on pensait être très caractéristiques de tel ou tel type, ne prenaient en fait aucune part à leur constitution. Par ailleurs, parmi les diverses hypothèses, également vraisemblables à priori, que s'était faites le chercheur, notre synthèse automatique a permis d'en retenir ou d'en rejeter un certain nombre ; ainsi il est apparu un nombre plus petit de types que celui attendu et dont certains n'étaient pas escomptés. Enfin, le degré de cohérence des différents types était tout à fait ignoré ; le spécialiste a été ainsi très surpris de la brutalité avec laquelle s'imposait le type "négatif".

3. *UN EXEMPLE SUR L'EVOLUTION DES ESPECES*

Cet exemple est tiré d'une étude présentée par Margaret Oakley Dayhoff dans *Scientific American*[1]. Le problème consiste à retrouver une classification phylogénique des êtres vivants à partir de la composition en acides aminés d'une certaine protéine la "cytochrome *c*", substance complexe, qu'on trouve chez les animaux et les plantes évoluées dans leurs mitochondries où elle joue un rôle biologique d'oxydation. Une telle protéine se présente sous la forme d'une séquence d'acides aminés de longueur donnée. Vingt séquences différentes de cette protéine, relatives à différents êtres vivants ont été identifiées et analysées. Chaque acide aminé étant représenté par une lettre, une protéine donnée se présente matériellement sous la forme d'un mot sur un alphabet fini. Tous les mots sont de même longueur à savoir 112.

Deux protéines sont considérées d'autant plus dissemblables que le nombre de mutations des acides aminés composants pour passer de l'une à l'autre est plus grand ; en d'autres termes, que le nombre minimum de lettres à changer dans des positions correspondantes pour passer d'un des mots à l'autre est plus grand.

L'arbre phylogénique à construire comprend des nœuds et des branches. Une classe d'êtres vivants indistinguables à une période donnée de leur évolution est représentée par un nœud dont sont issus deux ou plusieurs branches. A partir des séquences actuelles de la "cytochrome *c*", on peut déterminer, de proche en proche, une protéine type pour chacun des nœuds. A chaque branche est affectée une pondération sous la forme d'un nombre entier positif : le nombre de mutations pour transformer la protéine du nœud de l'une des extrémités de la branche en celle du nœud de l'autre extrémité. La pondération affectée à une chaîne de l'arbre est la somme des pondérations des branches constituantes.

La biologiste cherchera à déterminer l'arbre pour lequel la somme des pondérations des chaînes maximales est minimum. L'algorithme, proposé dans l'article cité, procède à partir de protéines connues actuellement, par tâtonnements successifs, envisageant un très grand nombre d'arbres possibles (sinon tous !) et ne retenant que le meilleur. Compte tenu de l'exploration exhaustive de l'algorithme à travers tous les arbres possibles , l'intérêt de ce dernier est pratiquement limité au problème particulier étudié, où la taille de la population à traiter est très modérée.

En fait, l'objet mathématique recherché n'est pas autre chose qu'une chaîne de partitions qu'on regardera à partir de la partition la moins fine jusqu'à celle la plus fine. Un niveau donné de la chaîne définira une

(1) "Computer Analysis of Protein Evolution" (Juillet 1969).

étape dans l'évolution des êtres vivants ; les classes de la partition de ce niveau correspondent aux noeuds de cette période. Du point de vue même du biologiste, deux protéines données dériveront d'un nœud (i.e. d'une classe) d'une période d'autant plus récente (i.e. d'un niveau de la chaîne d'autant plus élevé) qu'elles sont plus ressemblantes au sens précisé ci-dessus. Nous avons par conséquent envisagé de traiter le problème par notre méthode qui fournit une hiérarchie de classifications phénotypique.

La "cytochrome *c*", suite finie d'acides aminés, sera représentée par un vecteur logique de zéros et de uns qui se présente comme une suite de vecteurs logiques élémentaires ; chacun de ces derniers, étant associé à une position de la séquence de la manière suivante : Si, par exemple, $\mathfrak{A} = \{A , G , K , N\}$ est l'ensemble des acides aminés dont chacun est présent, à la i-ème position de la séquence, pour l'une au moins des vingt protéines ; le vecteur logique élémentaire associé comportera autant de composantes que $\mathfrak{A}$ d'éléments, ici quatre. La présence de A sera codée (1 , 0 , 0 , 0), celle de G (0 , 1 , 0 , 0), celle de K (0 , 0 , 1 , 0) et celle de N (0 , 0 , 0 , 1).

En ignorant les positions sans effet discriminatoire, où un seul acide aminé apparaît pour toutes les protéines, on se trouve ramené à ne considérer que 75 positions sur les 112 et à réaliser notre représentation au moyen de vecteurs à 246 composantes zéros ou uns pour lesquels le nombre de uns d'un même vecteur est égal à 75. Nos vingt objets, de la population sur laquelle il y a lieu de découvrir une hiérarchie de classifications, se trouvent ainsi situés dans l'hyperplan $\Sigma x_i = 75$ de l'hypercube $\{0,1\}^{246}$.

Dans ce cas tous les indices de similarité sont équivalents [cf. Chap. I § III] ; nous avons choisi le plus simple, à savoir s, pour établir la préordonnance sur l'ensemble des protéines. L'application de l'algorithme "lexicographique" a conduit aux résultats que nous exprimons dans l'annexe I sous le titre "exemple d'édition de l'arbre" et qui sont parfaitement concordants avec ceux de M.O. Dayhoff.

Par ailleurs nous avons étudié la classificabilité de notre population comparant la distribution observée $D(\omega)$ à celle de la statistique M dans l'hypothèse nulle N avec $T = 246$ et $k = 75$. On trouvera ci-après les deux graphiques correspondant à ces deux distributions. Signalons que dans un graphique donné, un segment horizontal représente une classe de la préordonnance ; sa longueur désigne la proportion des paires dans cette classe du préordre et sa hauteur la proportion de triplets dont la paire médiane et la paire supérieure enserrent la classe.

Il est manifeste, dans l'exemple étudié, que l'on est en présence d'une population très bien classifiable.

Nous avons enfin déterminé les niveaux les plus significatifs de la hiérarchie de classifications en calculant la valeur de la statistique

$$(|\omega \cap R \times S| - r.s/2)/\sqrt{r.s\,(f+1)/12}$$

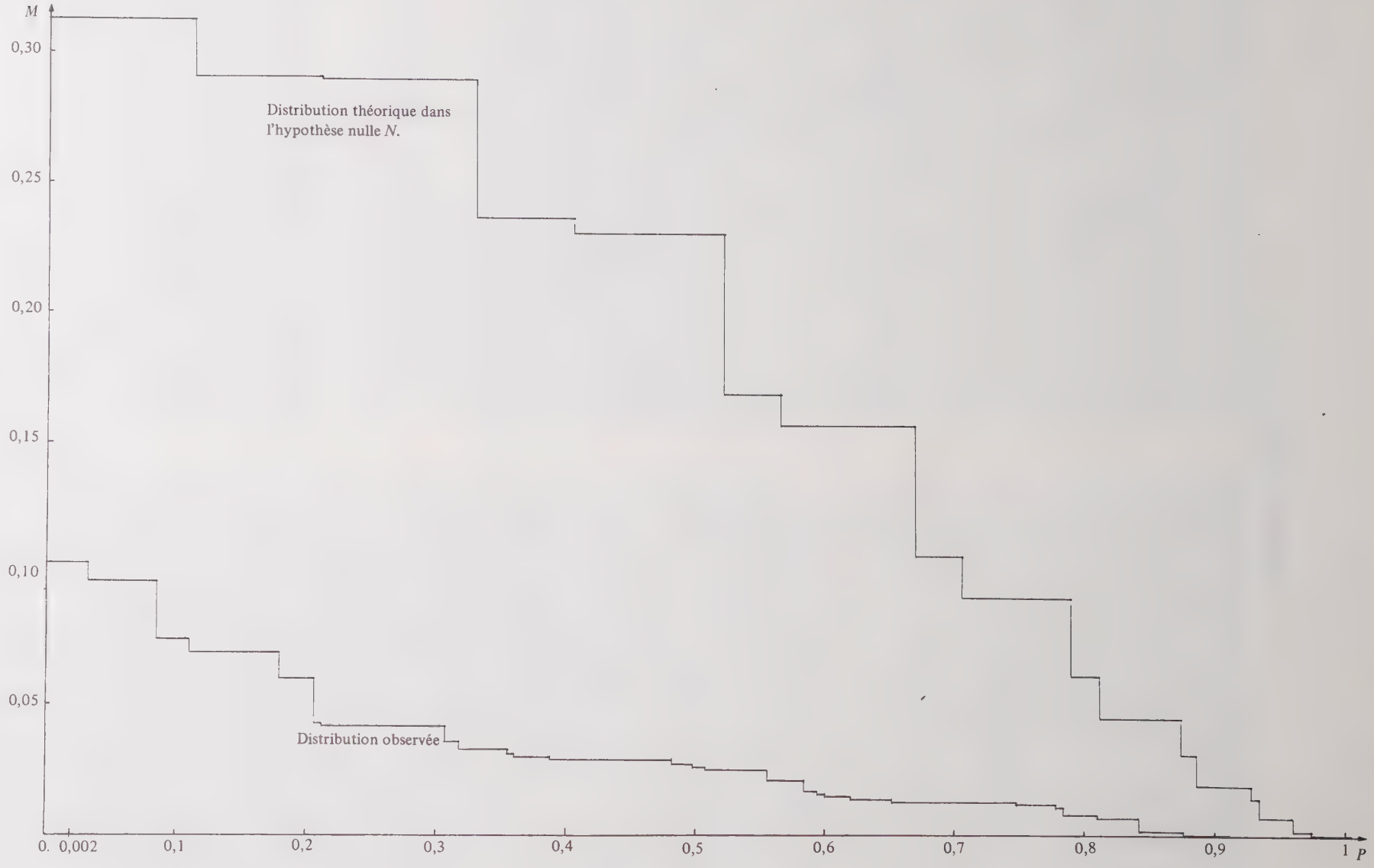
pour chacun des niveaux. [cf. Ch. II § VI.4]

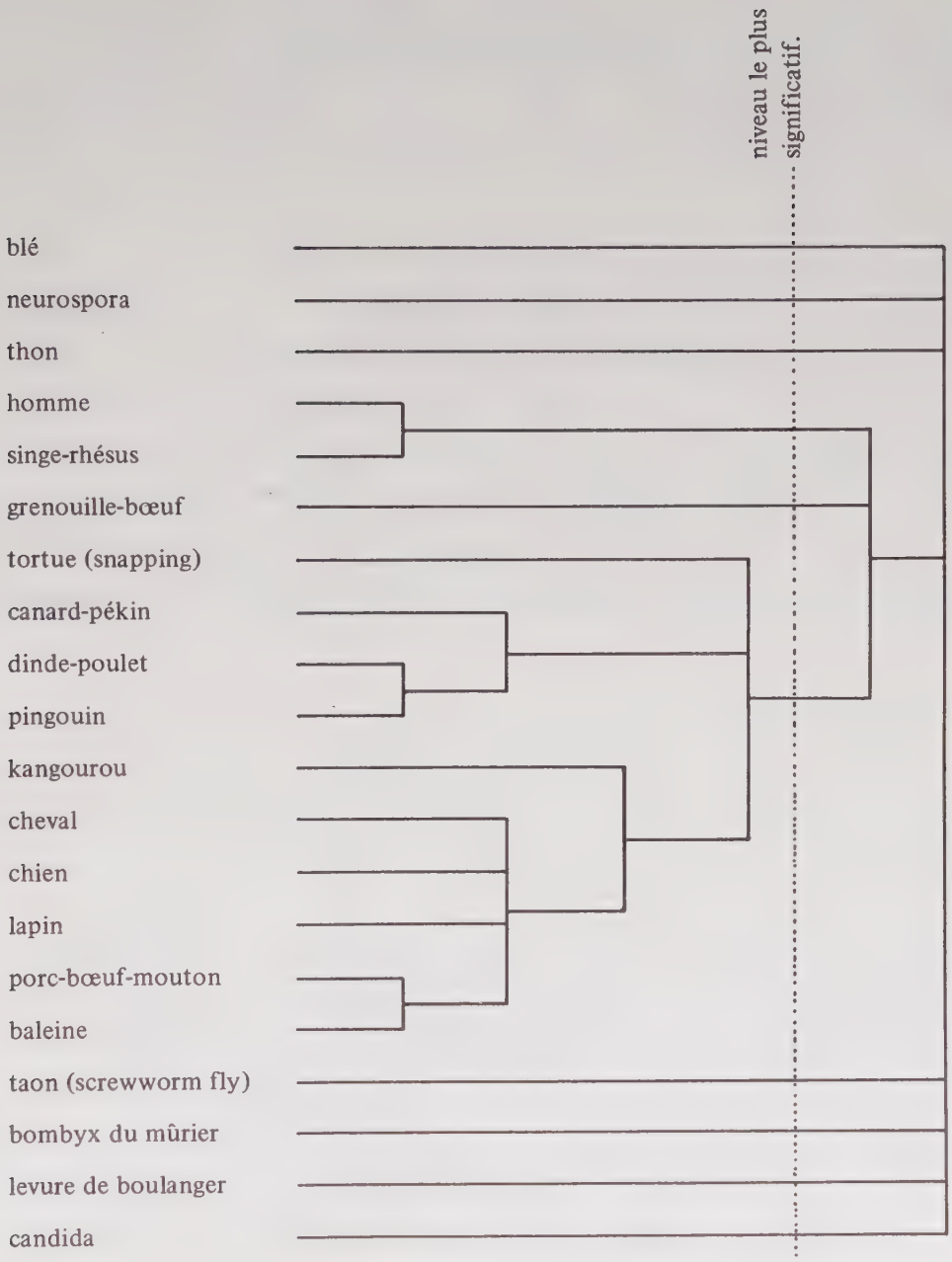
Le programme que nous avons établi pour cela comporte deux parties. La première lit la représentation polonaise de l'arbre établissant pour chaque paire le niveau de la hiérarchie où la paire se trouve réunie pour la première fois. La deuxième lit la suite des paires rangées selon l'ordonnance et charge, au fur et à mesure de cette lecture, un tableau définissant pour chaque niveau le nombre de paires déjà rencontrées et réunies à ce dernier ; tableau qui permet de remplir de proche en proche celui qui finalement contiendra pour chacun des niveaux la valeur de la statistique ci-dessus.

Les résultats pour notre étude sont exprimés dans le tableau ci-dessous dont la première ligne définit la suite décroissante des valeurs de la statistique et la seconde, les niveaux correspondants.

6.002	5.318	3.649	3.177	2.443	0.635	0.301	− 0.465	− 1.198	− 9.608
5	3	6	4	2	8	7	10	9	11

Compte tenu de la distribution de la statistique dans les hypothèses exprimées précédemment [cf. Ch. II § VI.4], nous retiendrons les niveaux 2, 3, 4, 5 et 6 pour obtenir l'arbre réduit suivant que nous laissons au naturaliste le soin d'apprécier.


M
0,30
0,25
0,20
0,15
0,10
0,05
Distribution théorique dans
l'hypothèse nulle N.
Distribution observée
0.
0,002
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1
P


niveau le plus significatif.
blé
neurospora
thon
homme
singe-rhésus
grenouille-bœuf
tortue (snapping)
canard-pékin
dinde-poulet
pingouin
kangourou
cheval
chien
lapin
porc-bœuf-mouton
baleine
taon (screwworm fly)
bombyx du mûrier
levure de boulanger
candida

BIBLIOGRAPHIE

BIBLIOGRAPHIE A

(affectée plus spécifiquement au chapitre 0)

[*a*] BIRKHOFF G. — Lattice Theory (A.M.S. 1967).

[*b*] BOURBAKI N. — (a) *Théorie des ensembles,* Chapitre 2 ; Paris, Hermann, (1966). (b) *Algèbre*, Chapitre IV, § 5 ; Paris, Hermann, (1959).

[*c*] KRASNER M. — Espaces ultramétriques et ultramatroïdes. Séminaire, Faculté des Sciences de Paris, année 1953-1954.

[*d*] RIORDAN J. — Introduction to combinatorial analysis ; New-York, Wiley, Chap. 6 ; 1958.

BIBLIOGRAPHIE B

(affectée aux chapitres 1 à 4)

[1] ADANSON M. — *Histoire naturelle du Sénegal, Coquillages* ; Paris, Bauche , 1957.

[2] BECKNER M. — The biological way of thought ; New-York, Columbia University Press, 1959.

[3] BENZECRI J.P. — (a) *Sur les algorithmes de classification. Problèmes et méthodes de la taxonomie.* Rennes, 1965, et Cours I.S.U.P. Paris 1965-1966. (b) *Ordre latéral entre lois de probabilité sur un ensemble ordonné ; application à l'Analyse Factorielle et aux critères de classification* (I.S.U.P.) Paris ; 1968.

[4] BONNER R.E. — On some clustering techniques. *I.B.M. Jour. Res. Devel.* 8, pp. 22-32, 1964.

[5] DAGNELIE P. — A propos des différentes méthodes de classification numérique ; *Revue de Statistique Appliquée, I.S.U.P.,* Vol. **XIV** n° 3, 1966.

[6] DANSER B.H. — A theory of systematics – *Bibl. Biotheor.*, **4**, pp. 113-180.

[7] DICE L.R. — Measures of the amount of ecologic association between species – *Ecology*, T. **26**, pp. 297-302, 1945.

[8] GILMOUR J.S.L. — A taxonomic problem ; *Nature*, **139**, pp. 1040-1042, 1937.

[9] GOWER J.C. — A comparaison of some methods of cluster analysis ; *Biometrics*, 23 pp. 623-637, 1967.

[10] JACKSON D.M. — Comparaison of Classifications ; *Proc. Colloquium of St-Andrews* 2-4 September 1968.

[11] JARDINE C.J., JARDINE N. and SIBSON R. — The structure and construction of taxonomic hierarchies, *Math. Biosc.*, Vol. 1, p. 173, 1967.

[12] JARDINE N. and SIBSON R. — *The measurement of dissimilarity,* (forthcoming paper).

[13] JOHNSON S.C. — Hierarchical clustering schemes, *Psychometrika*, Vol. 32, p. 241, 1967.

[14] KENDALL M.G. — *Rank Correlation Methods,* Charles Griffin, 1962.

[15] KRUSKAL J.B. — (a) Multidimensionnal scaling by optimising goodness of fit to a nonmetric hypothesis. (b) Nonmetric multidimensionnal scaling : a numerical method ; *Psychomatrika*, Vol. 29, p. 1 and p. 115 ; 1964.

[16] KULCZYNSKI S. — Die Pflanzenassoziationen der Pieninen ; *Bull. Inter. Acad. Pol. Sci. Lett. Cl. Sci. Math. Nat.*, (Sci. Nat.) pp. 57-203, suppl. 2 , 1927.

[17] LANCE G.N. and WILLIAMS W.T. — A general theory of classificatory sorting strategies I. Hierarchical systems. II Clustering systems ; *Computer Journal*, Vol. 9, p. 373 et Vol. 10 p. 271.

[18] LERMAN I.C. — (a) *Indice de Similarité et Préordonnance associée* ; rapport interne ; Centre de Calcul (M.S.H.), Paris , 1966 et communication à SEMINAIX ; 1967. (b) *Sur deux critères de la Classification* ; rapport interne ; Centre de Calcul (M.S.H.), Paris, 1967 et communication au "Colloquium of St-Andrews", 2-4 Septembre 1968. (c) *Analyse du problème de la recherche d'une hiérarchie de classification* ; rapport interne ; Centre de Calcul (M.S.H.), Paris, 1968.

[19] MAHALANOBIS P.C. — On the generalised distance in statistics;*Proc. Nat. Inst. Sci. India*, Vol. 2, p. 49.

[20] MOON J.W. — A note on approximating symmetric relations by equivalence relations ; *S.I.A.M. Jour. Appl. Math.*, 14, pp. 226-227, 1966.

[21] NEEDHAM R.M. — (a) *The theory of clumps II,* Cambridge Language Research Unit, M.L. 139, (Miméo), 1961. (b) Automatic classification in linguistics, *Statistician*, Vol. 17, p. 45, 1967.

[22] OCHIAI A. — Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions ; *Bull. Jap. Soc. Sci. Fish.* , T. 22, pp. 526-530 ,1957.

[23] ORLOCI L. — Definitions of structure in multivariate phytosociological samples ; *Vegetatio*, Vol. XV, Fasc. 5-6, pp. 281-291, 1968.

[24] PARKER RHODES A.F. — *Contributions to the Theory of Clumps.* Cambridge Language Res. Unit. 1961.

[25] PEARSON K. — On the coefficient of racial likeness ; *Biometrika*, 18, pp. 105-117.

[26] RÉGNIER S. — (a) *Sur quelques aspects mathématiques des problèmes de la classification automatique.* Rapport interne ; Centre de Calcul (M.S.H.) Paris ; et *I.C.C. Bull.*, Vol. 4, pp. 175-191, 1965. (b) Classification et Analyse des expressions plastiques non figuratives de malades mentaux. *Actes du Colloque*

International sur l'Informatique. Toulouse 7, 8, 9 Mars 1966. (c) 1. *Aspects statistiques*. 2. *Stabilité des opérateurs de Classification* ; rapport interne, Centre de Calcul (M.S.H.) Paris, 1966.

[27] ROGERS D.J. et TANIMOTO T.T. — A computer program for classifying plants ; *Science*, T. 132, pp. 1115-1118, 1960.

[27'] ROUX M. — *Algorithme pour construire une hiérarchie particulière*. Thèse de 3ème cycle —Stat. Math.—, Déc. 1968 ; Univ. Paris.

[28] RUSSEL P.F. et RAO C.R. — On habitat and association of species of anopheline larvae in south-eastern Madras ; *J. Malar. Inst. India* , T. 3, pp. 153-178,1940.

[29] SHEPARD R.N. — The analysis of proximities : scaling with an unknown distance function, I et II, *Psychometrika*, 1962.

[30] SNEATH P.H.A. — The application of computers to taxonomy ; *Jour. Gen. Microcrobiol.*, 17, pp. 201-226, 1957.

[31] SOKAL R.R. et MICHENER C.D. — A statistical method for evaluating systematic relationships ; *Univ. Kansas Sci. Bull.*, 38, pp. 1409-1438 , 1958.

[32] SOKAL R.R. et SNEATH P.H.A. — *Principles of numerical taxonomy*; San Francisco and London, Freeman and C°, 1963.

[33] SORENSEN T. — A method of establishing groups of equal amplitude in plant sociology based on similarity of species content ; *Konge Dan Vidensk. Selsk., Biol. Skr*, 5, pp. 1-34 , 1948.

[34] De la VEGA W.F. — Techniques de classification automatique utilisant un indice de ressemblance ; rapport interne, Centre de Calcul (M.S.H.), Paris, 1965 et *Revue Française de Sociologie*, Déc. 1967.

[35] VICQ d'AZYR F. — *Quadrupèdes ; Discours préliminaire ; Encyclopédie méthodique*, Vol. 2 ; Paris, Panckoucke, 1792.

[36] VO KHAC Kh. et NGHIEM Ph. T. — Etude sur les aspects théorique et pratique de la segmentation aux moindres carrés ; *Rev. Franç. d'Inform. et de Rech. Opér.*, n° 8, V. 1, pp. 77-90 , 2e année Mars/Avril, 1968.

[37] WARD J.H. — Hierarchic grouping to optimise an objective function ; *Journ. Amer. Stat. Assoc.*, Vol. 58, p. 236, 1963.

[38] WATANABE S. — Une explication mathématique du classement des objets In : *Information and prediction in science* (Ed. Dockx and Bernays), New-York and London, Academic Press.

[39] ZAHN C.T. — Approximating symmetric relations by equivalence relations, *S.I.A.M. Jour. Appl. Math.*, 12, pp. 840-847, 1964.

IMPRIMERIE LOUIS-JEAN
Publications scientifiques et littéraires
TYPO - OFFSET
05 - GAP - Téléphone 14-23 14-24
Dépôt légal 60 - 1970

Imprimé en France

Dépôt légal : 1741

N° Dossier : 69 34

DANS LA MEME COLLECTION

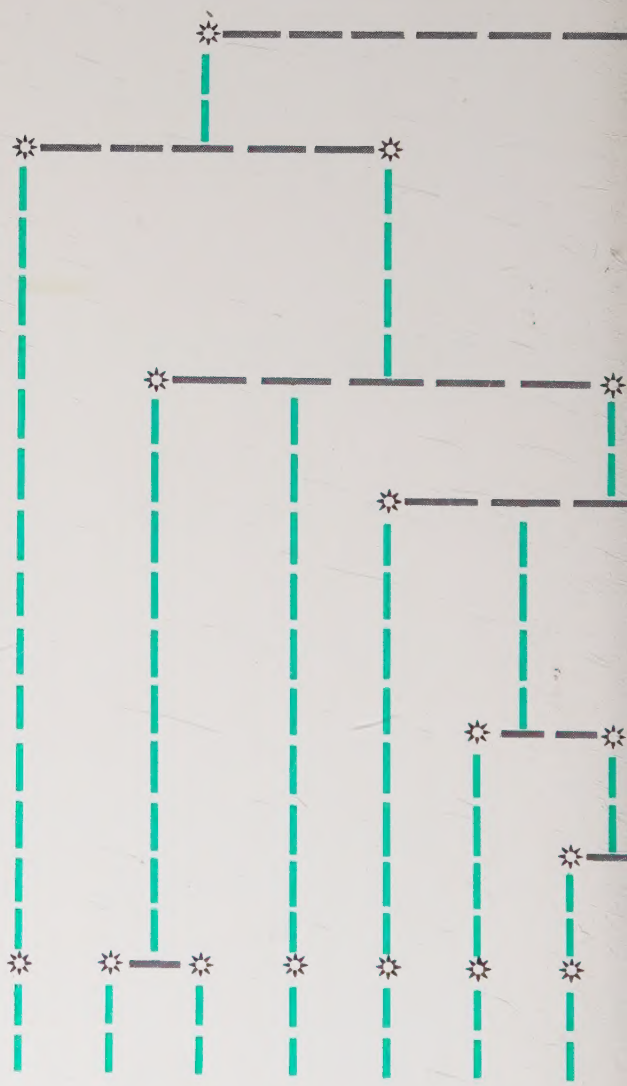
Parus :

ARSAC J., LENTIN A., NIVAT M. et NOLIN L. — Algol. Théorie et pratique.

GROSS M. et LENTIN A. — Notions sur les grammaires formelles.

FAURE R. — Eléments de la recherche opérationnelle. Avec la collaboration de *N.S. Guillot Le Garff* et de *M. Bloch.*

NOLIN L. — Formalisation des notions de machine et de programme.

Sous presse :

FAURE R. et HEURGON E. — Structures ordonnées et algèbres de Boole.